LAY, ALEXANDRA LESLIE, Ph.D. Accuracy, Stability, and Discriminant Validity Evidence: A Comparative Study of Exploratory and Confirmatory Methods for Assessing Statistical Dimensionality. (2020)
Directed by Dr. John T. Willse. 100 pp.

Defining the dimensionality of an educational or psychological assessment is a complicated task that requires great consideration given its potential for far reaching validity implications. This three-paper series of real data analyses aims to investigate how various methods for assessing dimensionality can work in tandem to strengthen claims regarding dimensional structure and whether there are administrative conditions, specifically related to sample size, in which some methods are more appropriate than others. The first paper provides a comparative analysis of the accuracy and precision of three exploratory dimensionality analyses, exploratory DETECT, common factor analysis (FA), and principal components analysis (PCA). The second paper investigates dimensionality at a finer grain size than the first, evaluating the stability of the item-to-factor mapping provided by exploratory DETECT. Finally, paper three begins to bridge the gap between exploratory and confirmatory analyses by employing the DETECT-defined factor structure with confirmatory DIMTEST and confirmatory factor analysis (CFA).

ACCURACY, STABILITY, AND DISCRIMINANT VALIDITY EVIDENCE: A COMPARATIVE STUDY OF EXPLORATORY AND CONFIRMATORY METHODS FOR ASSESSING STATISTICAL DIMENSIONALITY

by

Alexandra Leslie Lay

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2020

Approved by

_____
Committee Chair

APPROVAL PAGE

This dissertation, written by Alexandra Leslie Lay, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____

Date of Acceptance by Committee

_____

Date of Final Oral Examination

ACKNOWLEDGMENTS

I would like to thank my dissertation committee of Dr. John T. Willse, Dr. Terry Ackerman, Dr. Bob Henson, and Dr. Kyung Yong Kim for their support and iterative feedback throughout this process. Additionally, I would like to thank Dr. John T. Willse, Dr. Terry Ackerman, Dr. Micheline Chalhoub-Deville, and Dr. Ric Luecht for the many phases of mentorship and support they've provided me throughout my years at UNCG. Their continued guidance was integral to my development as a research professional. Lastly, I would like to acknowledge and thank Dr. Jamie Lynch at St. Norbert College for taking it upon himself to mentor a lost undergraduate student from outside his department. It was his constant encouragement and support that sparked my passion for research and set me on the path to pursuing a career in educational measurement.

I would also like to thank Bess Patton and Ramsey Cardwell for their friendship throughout the years. Bess has been an invaluable pillar of my support system and I am excited to continue our friendship and professional collaboration for years to come. Additionally, Ramsey has provided much needed comedic relief amidst of all the graduate school stress and for that I cannot thank him enough.

Finally, I would like to thank my personal village for their unwavering love and support throughout this process. To my husband, Shelby, thank you for your abundance of selflessness and sacrifices made to support me. Without you this degree would be a dream rather than a reality. To my daughter, Sofia, thank you for loving me through all

the time together lost and reminding me with your constant displays of fierce (and endearing) independence that this degree serves a greater purpose, this was for us.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The field of psychometrics has a long running history of evaluating the statistical dimensionality of instruments aimed at measuring psychological constructs, participant attitudes, and cognitive skill development and educational content mastery. Generally, available methods for assessing dimensionality fall into two broad categories: exploratory and confirmatory. Exploratory methods are typically employed when no substantive or otherwise theoretically defensible underlying factor structure exists, while confirmatory analyses are used to assess the level of statistical confidence one can have in previously defined factors (Reckase, 2009). Because a variety of methods for assessing dimensionality have existed for many years, a plethora of research exists overviewing their individual implementation, often with applicative examples.

However, some distinct gaps do exist in the literature, mostly in relation to evaluating the performance of dimensionality analyses. For instance, the current literature often fails to consider, or provide explicit instruction of how, the two general approaches to assessing dimensionality (i.e. exploratory and confirmatory) can work hand in hand to strengthen the conclusions being made regarding the factor structure of an assessment. Even when the potential for overlap is acknowledged (Schmitt, 2011), it is typically followed with ambiguity of direction for conducting such an analysis. Failing to extend dimensionality research to include a combination of exploratory and confirmatory

approaches is a missed opportunity to thoroughly flesh out any potential dimensional conclusions to be made, whether those conclusions prove to be in support of or refute a proposed factor structure.

Moreover, comparative analyses are generally limited to methods of the same classification (e.g. parametric, nonparametric, etc.), rather than extending evaluations of performance across classifications (Kogar, 2018). Again, as with failing to extend analyses to include both exploratory and confirmatory methods, not considering methods of different classifications is a missed opportunity to increase support for, or appropriately refute, a proposed factor structure. That is, implementing dimensionality analyses from multiple classifications helps reduce the entanglement of resulting conclusions to the method imposed, and as such, can increase confidence in those conclusions.

Additionally, the current literature on evaluating the analytic performance of dimensionality analyses is dominated by studies that employ simulated data. While simulation studies have proven effective for evaluating methodological performance across a wide range of technical combinations, the use of simulated data alone in the evaluation of dimensionality analyses is particularly problematic as results of these analyses are extremely context specific. That is, dimensionality is a function of the *interaction* between students and the assessment (Reckase, 2009); utilizing simulation studies to evaluate the performance of dimensionality analyses removes those resulting conclusions from the most integral piece to assessing dimensionality, the interaction.

References

Kogar, H. (2018). An Examination of Parametric and Nonparametric Dimensionality Assessment Methods with Exploratory and Confirmatory Mode. *Journal of Education and Learning*, *7*(3), 148-158.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79-112). Springer, New York, NY.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304-321.

CHAPTER II

OVERVIEW OF THREE PAPERS

In this dissertation, I investigated various dimensionality analyses with the goal of gaining a greater understanding of how these methods can work in tandem to strengthen claims regarding dimensionality and whether there are administrative conditions, specifically related to available sample size, in which some methods are more appropriate than others. Although bound together by the previously outlined purposes, each paper included in this dissertation is written to stand alone. As such, each paper contains its own specific research questions as well as introduction, methods, results and discussion sections.

The first paper provides a comparative analysis of the accuracy and precision of three exploratory dimensionality analyses, DETECT, parallel analysis (PA) implementing common factor analysis (FA), and PA implementing principal components analysis (PCA) in their systematic attempts at defining factor structure. For this paper one nonparametric and two parametric methods a employed and compared. The specific research questions addressed in the first paper were:

1. How does the accuracy of exploratory DETECT, PA implemented with common FA, and PA implemented with PCA change when evaluating dimensionality at different sample sizes

2. How does the precision of exploratory DETECT, PA implemented with common FA, and PA implemented with PCA change when evaluating dimensionality at different sample sizes?
3. What, if any, sample size specific recommendations can be made for choosing one method over another?

In the second paper, a more fine-grained investigation into exploratory dimensionality results was conducted, taking an in depth look at the performance of the factor definitions provided by exploratory DETECT. More specifically, the second paper introduces a factor congruence indictor to evaluate the stability of the exploratory DETECT factor definitions. The second paper aimed to explore the following research questions:

1. How does sample size affect the stability of factor definitions of the exploratory DETECT procedure?
2. How does the effect of sample size on factor definition stability compare to the results regarding stability and accuracy of the number of factors determined by the exploratory DETECT procedure in previous research?
3. Is there an advantage to evaluating performance of the exploratory DETECT procedure at a smaller grainsize (i.e. factor composition level) or is the more parsimonious approach to evaluating the number of factors sufficient?

Finally, paper three begins to bridge the gap between exploratory and confirmatory analyses. Building on the factor structure defined in the first two papers, the third paper employs a variety of confirmatory analyses (i.e. confirmatory DIMTEST and

confirmatory factor analysis (CFA)) for the purpose of building discriminant validity evidence of the exploratory structure. As with the first two papers, the third also considers how sample size may impact the performance of the dimensionality analyses. The final set of research questions include:

1. Do confirmatory results from CFA and DIMTEST tend to support the factor structure provided by exploratory DETECT?
2. Did one confirmatory technique recover the DETECT factor structure better than the other?
3. What is the effect of sample size on the effectiveness of using these techniques to build discriminant validity evidence for the DETECT-specified factor structure?

CHAPTER III

SIGNIFICANCE

Dimensionality analyses aren't new to the field of psychometrics, nor are the methods to be utilized in this dissertation. However, there has been a lag in the academic research since their induction as educational testing organizations increased their efforts towards standardization and it became common practice, to the best of their efforts, to build unidimensional tests. As tests are seldom truly unidimensional (Ackerman, Gierl, & Walker, 2003; Walker, Azen, & Schmitt, 2006) thorough investigations into their underlying factor structure are warranted, and one may even argue, more pertinent in the context of standardized assessments. That is, considering most high-stakes assessments (e.g. state accountability assessments, college entrance exams, etc.) make the claim of unidimensionality, confirming this claim statistically prior to making high-stakes decisions is of utmost importance.

Additionally, the current climate in educational measurement is rapidly moving in the opposite direction of standardization towards a more fluid conceptualization defined by ever-evolving, technologically complex assessments (DiCerbo, 2014; Ventura & Shute, 2013; Drasgow, 2015; Lay, Patton, & Chalhoub-Deville, 2017). More specifically, as the field of psychometrics catapults towards more dynamic learning and assessment environments such as game-based assessments and interactive, online instruction, the potential for these educational contexts to elicit multidimensionality becomes high while

confidence in their "true" underlying construct or factor structure remains low (Lay, Patton, & Chalhoub-Deville, 2017). The shift towards these complex assessment types necessitate that increasing energy be devoted to the evaluation and verification of the underlying factor structure of a given assessment.

In both approaches to developing educational instruments, standardization and fluidity, the obligation to evaluate dimensionality is abundantly clear. However, the field is still lacking clear direction into how to adequately conduct thorough dimensionality investigations, prior to score analysis and reporting, when the underlying factor structure is unknown. This dissertation's purpose is to move the field closer to developing such a methodological road map that includes a wide range of dimensionality analysis options and provides examples of context specific differences in performance of the overviewed methods. While recognizing that statistical dimensionality is the function of the interaction between relevant context-specific examinee characteristics and the assessment at hand (Reckase, 2009), the contextual focus will remain on sample size throughout the papers to limit the scope of variability between iterations of analysis as I attempt to parse out differences in the performance of the methods.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-51.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society*, *17*(1), 17-28.

Drasgow, F. (2015). *Technology and testing: Improving educational and psychological measurement*. Routledge.

Lay, A., Patton, E., & Chalhoub-Deville, M. (2017). A case for the use of the ability-in language user-in context orientation in game-based assessment. *Language Testing in Asia*, *7*(1), 16.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79-112). Springer, New York, NY.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568-2572.

Walker, C. M., Azen, R., & Schmitt, T. (2006). Statistical versus substantive dimensionality: The effect of distributional differences on dimensionality assessment using DIMTEST. *Educational and Psychological Measurement*, *66*(5), 721-738.

CHAPTER IV

PAPER 1: A COMPARATIVE STUDY OF THE ACCURACY AND PRECISION OF EXPLORATORY DIMENSIONALITY ANALYSES

Introduction

A vast variety of methods exist for estimating and assessing statistical or psychometric test dimensionality. At the broadest level, these methods may be categorized as exploratory or confirmatory analyses. Typically, exploratory analyses allow the data to define the dimensional structure of the test, searching for a definitive partitioning of factors when statistical multidimensionality exists. Conversely, confirmatory analyses serve as a statistical check on a previously defined structure within the data. By this categorization, examples of exploratory analyses include exploratory factor analysis (EFA), principle components analysis (PCA), and exploratory DETECT while confirmatory factor analysis (CFA) and confirmatory DIMTEST are examples of confirmatory analyses. Additionally, Jang and Roussos (2007) offer an alternative organization of theses analyses, those that "attempt some degree of full dimensionality assessment (number of dimensions and which items measure which dimensions)" (e.g. EFA, exploratory DETECT, etc.) or those that "merely attempt to assess lack of unidimensionality" (e.g. confirmatory and exploratory DIMTEST, etc.) (pp. 5). Due to the variety of terminology used to describe the clustering of items within a set of examinee response data (e.g. clusters, factors, dimensions, etc.) across the various classifications of dimensionality analyses, all dimensionality based subsets of items will

be referred to as factors for the remainder of this study in the interest of clarity and consistency.

Irrespective of which method is selected or the purpose of analysis, central to the task of evaluating statistical dimensionality is determining the dimensional structure of the test. That is, the usefulness of confirmatory analyses relies on the accuracy of the defined dimensional structure, and while exploratory analyses are intended to help inform the structure, there is much debate as to how to correctly utilize exploratory results. Using the incorrect number of factors can have serious implications, as noted in Franklin, Gibson, Robertson, Pohlmann, and Fralish, (1995):

> An incorrect choice may lead to the underextraction of factors (i.e. loss of information), but usually results in overextraction (i.e. inclusion of spurious factors). Overextraction of factors attaches meaning to noise and results in the interpretation of random variation in the data, thus affecting subsequent analyses or factor rotations (Zwick & Velicer, 1986), (p.100).

Due to the multitude of potential downstream impacts of incorrect factor extraction, it is important to use more statistically sound methods for determining factor extraction or retainment than the historically relied upon 'rules of thumb', commonly implemented via scree plots (Singh & West, 1971; Franklin et al., 1995; Watkins 2006).

Before continuing the discussion of factor and factor extraction techniques, the delineation of, and conversely the relationship between, EFA and PCA will necessarily be fleshed out for the readers. At the conceptual level, PCA may be referred to as simply a data reduction technique aimed at accounting for as much variance as possible with as few variables as possible, while the goal of EFA is to reveal the latent structure (i.e.

commons factors) that causes the observed variables to covary with one another (Fabrigar & Wegener, 2011; Osborne, Costello, & Kellow, 2008). Mathematically, this fundamental difference is explicit in the way each analysis considers sources of variance. That is, in PCA "components are calculated using all of the variance of the manifest variables, and all of that variance appears in the solution" (Ford, MacCallum, & Tait, 1986, pp. 88) while in EFA "the shared variance of a variable is partitioned from its unique variance and error variance to reveal the underlying factor structure; only shared variance appears in the solution" (Osborne, Costello, & Kellow, 2008, pp. 2).

Considering the aforementioned distinction, PCA could clearly be referred to as a standalone methodology, however, it is more often regarded as a branch or subset of EFA (Fabrigar & Wegener, 2011; Hayton, Allen, & Scarpello, 2004). That is, despite their distinctions, both methods are exploratory in nature and they both aim to identify the underlying factor structure of the data. This perspective will be used throughout the remainder of this study as PCA and EFA are both considered factor extraction techniques. To refrain from confounding the two as a single technique, while still preserving their relationship as exploratory methods, they will hereafter be referred to as PCA and common factor analysis (FA), both falling under the umbrella of EFA models.

One statistical method for determining the number of factors to extract is Horn's (1965) Parallel Analysis (PA). Support for the use of PA is widely documented in the literature. Watkins (2006) asserts that PA is one of only two consistent methods for determining factors to extract, with the other being the Minimum Average Partial (MAP; Velicer, 1976). Moreover, while Hayton, Allen, and Scarpello (2004) note that there is

little consensus on which method to use, PA emerges as one of the most accurate, while Thompson and Daniel (1996) exclusively recommend using PA for factor extraction. Finally, a study by Henson and Roberts (2006) concluded that, of the methods reviewed, PA was the most accurate procedure for factor retention.

PA is a sample-based adaptation of Kaiser's (1960) rule to retain factors with eigenvalues greater than 1, intended to resolve the issue of factor indeterminacy (Franklin et al., 1995). That is, in PA the unrotated eigenvalues from the actual data are compared to those from a random correlation matrix of identical size (Franklin et al., 1995; Watkins, 2006). Typically, factors are retained when the eigenvalues from the actual data exceed those from the random data (Franklin et al.,1995; Watkins, 2006; Williams, Onsman, & Brown, 2010). It may be implemented with either common FA or PCA. Generally, Kaiser's rule (1960) is preserved during use with PCA. That is, if the variables are perfectly uncorrelated, as modeled by the random eigenvalues, they will explain the same amount of standardized variance at a value of 1 (Franklin et al., 1995; Dinno, 2018; Dinno, 2014). Therefore, retention with Horn's PA (1965) is theoretically dependent on both criteria: (1) Exceeding a value of 1, and (2) Exceeding the value of the corresponding random eigenvalue. This process is extremely similar with use of common FA, with the exception of a slightly different adjustment, which reduces the criterion value to 0 in accordance with the theoretical justification that if the variables are perfectly uncorrelated, there should be no common variance (Dinno, 2014). A greater technical description of this process is offered in the methods section.

Another method used to determine dimensional structure is exploratory DETECT (Kim, 1994; Zhang & Stout, 1999). A clear distinction between exploratory DETECT and EFA factor retention techniques will be maintained as DETECT differs fundamentally in multiple ways. First, unlike most EFA methods, DETECT is a nonparametric conditional covariance-based procedure. Second, DETECT assumes approximate simple structure in the data, and therefore attempts to map each item to the factor it predominately measures (Deng, Han, & Hambleton, 2012). When determining dimensional structure, "DETECT searches for a partitioning of the items into clusters such that the conditional covariances between items from the same factor are positive and the conditional covariances between items from different factors are negative" (Jang & Roussos, 2007, pp. 6-7). Once DETECT reaches the optimum factor partition based on within and between factor conditional covariances, it outputs the resulting number or factors with accompanying item-to-factor mapping, and a variety of statistics that evaluate the appropriateness of the given dimensional output. As with PA analyses, the more technical aspects of DETECT will be overviewed in greater detail later in the paper.

Like PA, the use of DETECT is widely advocated for and supported in the literature. One reason DETECT is advocated for over parametric EFA methods is that being nonparametric makes it comparatively less computationally intense, requiring only simplistic statistical computations instead of complex algorithms (Jang & Roussos, 2007; Roussos, Stout, & Marden, 1998, Roussos & Ozbek, 2006). Moreover, Roussos and Ozbek (2006) assert that "nonparametric techniques have become increasingly popular because they avoid strong parametric modeling assumptions while still adhering to the

fundamental principles of item response theory (IRT)" (p. 214). However, possibly the most practical advantage to using DETECT relates to its function of modeling approximate simple structure. That is, if approximate simple structure holds, sorting items into unique factors can aid in substantive interpretation of those factors, or as Zhang and Stout (1996) claim, the resulting factors are "referred to as the *correct* [or] *dimensionally-based*" and they are considered "substantively meaningful and dimensionally separate" (Zhang & Stout, 1996, pp. 214; emphasis original).

Purpose

While the current literature on statistical test dimensionality includes thorough overviews of the available methods and applicative examples, it fails to provide an in-depth comparative analysis of a variety of those methods that includes statistically defensible recommendations for their use based on practical considerations such as available sample size. The purposes of this study were to (1) critically analyze a variety of methods for assessing statistical dimensionality by comparing them on metrics regarding their accuracy and precision, and (2) provide appropriate use-case specific recommendations based on the comparative results.

Research Questions

1. How does the accuracy of exploratory DETECT, PA implemented with common FA, and PA implemented with PCA change when evaluating dimensionality at different sample sizes?

2. How does the precision of exploratory DETECT, PA implemented with common FA, and PA implemented with PCA change when evaluating dimensionality at different sample sizes?
3. What, if any, sample size specific recommendations can be made for choosing one method over another?

Methods

Data Sources

One form of a large-scale standardized math test was used in this study. The test contains 60 items spanning eight content areas and three cognitive skill domains. Response data is available for 25,000 examinees, but additional samples of varying size were composed for analysis.

Analyses

As the purpose of this study was to evaluate the accuracy and precision of various dimensionality analyses by sample size, additional samples were composed prior to analysis. Bootstrap resampling (i.e. sampling *with* replacement) was employed to develop each additional sample. There were five sets of samples created to represent very small (*n=250*), small (*n= 500*), intermediate (*n=2,500)*, moderate (*n=10,000*) and large (*n=25,000*) sample sizes, each set containing 50 samples of the same size.

Three different exploratory dimensionality procedures were employed to evaluate the dimensional structure of the data and compare their relative degrees of accuracy and precision by sample size. For the purpose of this study, accuracy was defined as the degree of deviation of the dimensional results from the "true" dimensional structure of

the data. Relative truth in this case was defined by baseline dimensionality results obtained from analyzing the original sample of 25,000 examinees. Relative truth is also method specific, results in three "true" factor solutions in all. Moreover, precision is be defined as the level of variation in dimensionality results between samples within a sample size for each method.

The three distinct approaches for determining dimensionality of the examinee response data that were included in the analysis are PCA, common FA, and exploratory DETECT. Both PCA and common FA was implemented via PA. As previously mentioned, the major difference between the two exploratory analyses to be implemented with PA lies in how they conceptualize the variance between variables. Unlike PCA, common FA attempts to separate unique and shared variance of each variable prior to analysis. That is, the difference in the partitioning of variance between the two methods is directly reflected in the composition of the correlation matrix, where PCA utilizes the original correlation matrix with the diagonal elements all equal to one (i.e. implying that all variance is common), and common FA replaces the diagonal elements in the correlation matrix with the squared multiple correlations to represent the commonalities among variables separated from their unique variance (Rencher, 2002).

The *paran* package (Dinno, 2018) was used for execution of the PA in the programming environment R. The evaluation criteria for the PA package by Dinno (2018) deviates slightly from the traditional PA factor retention criteria, applying an adjustment to the real data eigenvalues to account for sampling error and least squares bias as described by Horn (1965) and retaining adjusted factors greater than 1 for PCA or

0 for common FA. That is, the procedure is organized as follows (Dinno, 2014; Dinno, 2018):

1. Conduct a PA implementing PCA or common FA with the real data to obtain the real eigenvalues: λ_q, where q represents the eigenvalue index, in which they are ordered in magnitude from largest to smallest, ranging from 1 to the number of total variables.
2. Conduct a PA implementing PCA or common FA (whichever was used in step 1) on a matrix, of an identical size to the real data used, of uncorrelated random values to obtain the random eigenvalues: λ_q^r, where r is an indicator of the random data matrix used for computation.
3. Repeat step 2 a specified amount of times (the default value for the *paran* package is 30**P*, where *P* is the number of variables contained in the real data) and obtain the average random eigenvalues: $\bar{\lambda}_q^r$.
4. Use the average random eigenvalues to compute the adjusted values from the real values. For PCA:

$$\lambda_q^{adj} = \lambda_q - (\bar{\lambda}_q^r - 1). \quad (4)$$

For common FA:

$$\lambda_q^{adj} = \lambda_q - \bar{\lambda}_q^r. \quad (5)$$

The final approach to evaluating dimensionality was the nonparametric, conditional covariance-based statistical procedure, exploratory DETECT (Kim, 1994; Zhang & Stout, 1999). DETECT provides a host of statistics to determine the magnitude of multidimensionality in the data, with the primary statistic being the DETECT value. The DETECT value reports the mean of all item-pair conditional covariances as a measure of the magnitude of multidimensionality (Roussos & Ozbek, 2006), and is calculated as follows:

$$D(P, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \le i \le ln} \delta_{i,l} E[cov(U_i, U_l | \Theta_{TT})], \quad (6)$$

where *P* represents the given partition of items in the assessment, *n* is the number of items on the assessment, *i and l* are any two items on the assessment with scores U_i and U_l, Θ_{TT} represents "the unidimensional latent variable 'best measured' by the total test [number correct] score", and $\delta_{i,l} = 1$ if item i and l are in the same factor and -1 if they are in different factors (Stout et al., 1996, p. 333). DETECT index values less than or equal to .1 indicate probable unidimesionality while values greater than 1.0 suggest the presence of multidimensionality (Stout, 1989; Stout et al., 1996). For the DETECT index estimate to reach its maximum possible value, all within-factor item pair covariances must be positive while all between factor item pair covariances are negative (Stout et al., 1996). The maximum DETECT value may be obtained by:

$$D^*(P, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \le i \le ln} |E[cov(U_i, U_l | \Theta_{TT})]|, \quad (7)$$

equation notation follows that of the previously defined DETECT value.

Additionally, DETECT provides the ratio r. The ratio r compares the computed DETECT value to its maximum possible value. The ratio is computed by

$$r = \frac{D(P,\Theta_{TT})}{D^*(P,\Theta_{TT})}. \quad (8)$$

The higher the ratio the more confidently multidimensionality can be concluded. The maximum value of 1 indicates simple structure, while .8 suggests approximate simple structure (Svetina & Levy, 2012). DETECT was employed from the *sirt* (Robitzsch, 2019) package in the R programming environment.

As previously mentioned, prior to evaluating dimensionality of the samples, baseline figures were determined. That is, the dimensional structure of the full sample of examinee response data (n= 25,000) were evaluated using all three methods. While some variation in results is expected to occur between methods, the results from each served as the "true" dimensional structure under each respective method.

After "true" dimensional structure is defined, all three approaches to assessing dimensionality were employed with each set of samples. Results from each group of analyses were then evaluated for accuracy and precision. As previously noted, accuracy is defined as the degree and direction of deviation from the "true" dimensional structure, and as such was evaluated relative to the number of factors in the baseline results for each set of samples by each dimensionality approach.

To represent the degree of deviation, mean absolute differences (MAD) was calculated at each sample size for each method with lower values representing higher accuracy. The MAD was calculated as follows:

$$MAD = \frac{\sum_{i=1}^{ns} |(S_{i,f} - T_f)|}{ns}, \tag{9}$$

where T_f represents the "true" number of factors, $S_{i,f}$ represents the number factors from a given sample, i is any given sample, and ns represents the number of samples. Lower MAD values indicate a lower degree of deviation from relative truth, with a value of zero representing a perfect match between the two.

Additionally, to evaluate the direction of deviation the mean error (ME) values are provided, defined here as simply the average differences between the true number of factors and the number extracted from a given sample. The ME was calculated similarly to the MAD:

$$ME = \frac{\sum_{i=1}^{ns} (S_{i,f} - T_f)}{ns}, \tag{10}$$

with notation defined in accordance with the MAD equation. ME values closer to zero represent greater accuracy. Precision is defined as the average standard deviation within set of samples for each approach at each sample size, which higher average standard deviations representing lower levels of precision.

Results

Data

As previously noted, the original dataset consisted of responses from 25,000 examinees on 60 items. The mean total score from the original dataset was 30.87 (*sd=11.65*), indicating that the average student answered just over 50% of the questions correct. Additionally, the median total score was 30, slightly lower than the mean, and visual inspection of the density distribution of total scores indicated that the data is very slightly positively skewed. Consistency in sampling distributions was also inspected, with samples increasing in consistency as sample size increases, but remaining within reasonable deviation at the lower sample sizes as well. Figure 4.1 provides a visual of the sample consistency results.

Defining Relative Truth

An initial analysis was completed employing each of the three methods with the original 25,000 observations. The goal of this analysis was to establish relative "truth" for each method, defined as the resulting number of factors in the unaltered data by method. One key difference of the exploratory DETECT procedure from the parallel analyses is that it requires a user-defined maximum number of factors to estimate. This user-defined value can range from two to the total number of items in the data. For the initial analysis the max number of factors to be estimated was freed to equal 60, the total number of items in the data. Initial results of the exploratory DETECT procedure and the common FA were consistent, indicating there were 15 factors in the data. The PCA resulted in substantially fewer factors than the DETECT and common FA analyses with a total of 5

factors. After establishing relative truth, the analyses were run on all 50 samples of each size. All subsequent DETECT analyses employed a maximum number of factors to estimate of 20 based on the baseline results of 15. See Figure 4.2 for boxplot results of all analyses.

Accuracy

When comparing sample results of the DETECT analysis to the relative true number of factors, accuracy levels out at the sample sizes of 2,500 and above. That is, the range of the MAD and ME values is .26 (*3.74 – 4.00*) and .10 (*-3.66 – -3.76*) respectively for sample sizes between 2,500 and 25,000, while the values are notably higher for the samples of lower size (*MAD = 5.78 – 7.42; ME =-5.78 – -7.42*). The samples of size 250, on average, deviated the most from the relative truth with a MAD value of 7.42 and a ME value of -7.42. Unsurprisingly, the samples of size 25,000, on average, deviated the least from the true number of factors with a MAD value of 3.74 and a ME value of -3.66. ME values were consistently negative across all sample sizes indicating that when deviating from relative truth, the DETECT analysis consistently under extracted the number of factors.

Similarly, sample results from the PCA indicate that the measures of accuracy become consistently minimal at samples of size 10,000 and above. More specifically, the range of MAD and ME values both reduce to .18 between samples of sizes 10,000 and 25,000, with higher values at the lower sample sizes. Consistent with the DETECT analysis, the samples of size 250 deviated, on average, the most from the relative PCA truth with a MAD value of 2.8 and a ME value of -2.8. Diverging slightly from the

pattern of DETECT results, the samples of size 10,000 deviated the least from relative PCA truth with a MAD value of 0.04 and a ME value of -0.04. ME values were consistently negative for samples of sizes 250, 500, and 2,500, and exactly the inverse of their corresponding MAD values suggesting that when the PCA results deviated from relative truth at those sample sizes, they were always under extracted compared to relative truth. Across samples of size 10,000 and 25,000 the ME values exactly equaled their MAD values, indicating that when the results deviated from relative truth, they were always over extracted.

Finally, unlike the DETECT and the PCA results, the common FA sample results did not indicate a point at which stability in the number of factors estimated was likely to occur. More specifically, as sample size increased, the number of factors estimated by the common FA also continued to increase, eventually exceeding, on average, the "true" number of factors. However, results were consistent with the other methods regarding points of minimum and maximum deviation. As with the DETECT and PCA results, the common FA deviated the most from its relative truth with samples of size 250, resulting in a MAD value of 11.64 and a ME value of -11.64. Additionally, like the PCA analysis, the common FA deviated the least from relative truth with samples of size 10,000, resulting in a MAD value of 0.76 and a ME value of -0.16. ME values for sample sizes of 250, 500, 2,500, and 10,000, were all negative, indicating that when the sample results deviated from the true number of factors, they were mostly under extracted. However, for samples of size 25,000 the ME value is positive, as well as exactly equal to the

corresponding MAD value, indicating that when deviating from relative truth the results were consistently over extracted.

When comparing average accuracy measures across methods and all sample sizes, the common FA analysis emerged as the least accurate with a mean MAD of 5.792. Conversely, with a mean MAD value of 1.232, the PCA was the most accurate method of the three. When evaluating accuracy across methods, within sample size, samples of size 250 were least accurate with a mean MAD of 7.287. Additionally, samples of size 10,000 appeared to be most accurate with a mean MAD of 1.600.

Precision

To evaluate the degree of precision by method and sample size, standard deviations were computed. That is, the less variability within a set of sample results, the more precise the method can be concluded to be at that sample size. For the DETECT procedure, the results appeared to be most precise for samples of size 250, with a standard deviation of 1.486, and least precise for samples of size 10,000 with a standard deviation of 2.378. Results differed notably for the PCA, where samples of size 10,000 were most precise, with a standard deviation of 0.198, and samples of size 500 were least precise with a standard deviation of 0.606. Finally, for the common FA samples of size 25,000 were the most precise, resulting in a standard deviation of 0.935, while samples of size 2,500 were the least precise with a standard deviation of 1.753. Across sample sizes, the DETECT procedure was the least precise with an average standard deviation of 1.999, and the PCA was the most precise with an average standard deviation of 0.418. Across methods, samples of size 250 were the most precise with an average standard

deviation of 0.951, and samples of size 2,500 were the least precise with an average standard deviation of 1.434. Full aggregated accuracy and precision results can be found in Table 4.1.

Bias

While not a main focus of this study, it is worth noting that for both the PCA and common FA analyses, the amount of bias in which the eigenvalues were adjusted for, continuously decreased by sample size. For the PCA, the maximum mean bias value was 1.057 (*sd=.078*) at a sample size of 250, while the minimum mean bias value was .082 (*sd=.001)* at a sample size of 25,000. Simiarly, for the common FA, the maximum mean bias value was 1.258 (*sd=.039*) at a sample size of 250, while the minimum mean bias value was .062 (*sd=.001)* at a sample size of 25,000. Full bias results for both analyses can be found in Figure 4.3.

DETECT Statistics

Finally, as with the bias results, the DETECT statistics were not a major focus of this study. However, it would be remiss to exclude the ratio R results, as they offer an additional metric potentially related to the accuracy of the DETECT procedure. As a reminder, the ratio R values serve as an indicator of the similarity between the calculated DETECT value for a given sample and the maximum possible value, where 1 indicates a perfect match between the two. The ratio R results from this study continued to increase as sample size increased. That is, as sample size increased, and the number of factors extracted increased, the similarity between the estimated DETECT values and the maximum possible value increased. It is unclear whether (or how much of) this result can

be attributed to the increase in the accuracy of the procedure, rather than being just a product of the sample size itself, but it is worth noting, nonetheless. The minimum mean ratio R value was .432 (*sd=.014)* at a sample size of 250, while the maximum mean value was .843 (*sd=.010*) at a sample size of 25,000. The ratio R results can be found in Figure 4.4.

Discussion

Main Findings

Each exploratory analysis offered a unique set of results. The DETECT procedure was by far the least precise across the board in terms of variability within sample sizes and remained comparatively imprecise across sample sizes. Additionally, the sample results seldom matched the relative truth, instead the number of factors were consistently under extracted. However, there were a few points of positivity for the DETECT procedure. For one, the results leveled out after reaching the samples of size 2,500, indicating that at least within the range of 2,500-25,000, one could expect results of the analysis to be very similar and may suggest that more moderate sample sizes are sufficient for analysis. Additionally, the range of the mode factors by sample size was relatively small with a minimum of 7 and a maximum of 11.

By the accuracy and precision standards outlined in this study, the PCA significantly outperformed the other two methods. That is, there was minimal variability between results within sample sizes, suggesting the analysis is extremely precise at any of the given sample size. Moreover, at samples of size 10,000 and above, the PCA accurately extracted the correct number of factors relative to “truth” in the vast majority

of analyses. While the PCA has the DETECT procedure beat in terms of accuracy and precision, the two sets of results did have some similarities. Like the DETECT procedure, the PCA had a similar leveling effect at the sample sizes 10,000 and 25,000. Additionally, the PCA also had a very small range for the mode number of estimated factors across sample sizes with a minimum of 2 and a maximum of 5.

The results of the common FA were overall more precise within sample sizes than the DETECT analysis, but less precise than the PCA. While the common FA did reach a sample size at which it extracted the correct number of factors most of the time (*n=10,000*) as with the PCA, it also diverged from both other analyses by reaching a sample size at which it routinely exceeded the correct number of factors to extract as well (*n=25,000*). Additionally, unlike the other two methods, the common FA never reached a point at which the results leveled out. Finally, the range of the mode number of factors extracted by sample size was the largest of the three methods with a minimum of 3 and a maximum of 18.

While the main purposes of this study were to evaluate the accuracy and precision of these methods across sample sizes, the variety of results obtained between methods makes it difficult to develop any cohesive conclusions supported by all three. Moreover, with regards to comparative superiority of the methods, while the PCA appeared to outperform the DETECT and common FA analyses, a greater discussion centered around the purpose of analysis is warranted (and subsequently provided in the following section). All results considered, neither a single superior method, nor sample size-based recommendations, emerged from this study.

Considering Methodological Purpose

While all exploratory methods employed in this study were poised as approaches for determining the number of factors to extract, it is necessary to consider their more specific purposes when interpreting results. More specifically, initial analysis to determine the "true" number of factors relative to each method using the original sample of 25000 revealed that that the DETECT procedure and common FA extracted the same number of factors, while the PCA resulted in substantially fewer. The DETECT procedure and common FA methods are intended to model the latent structure of the data, while by contrast the PCA is strictly a data reduction technique. This difference may provide an additional facet of clarity to the discrepancy in results between methods, specifically with how distinctly different PCA results were from the other two methods. Moreover, the results may provide additional support for researchers to exercise great caution when determining which method to use for factor extraction, putting appropriate weight on the purpose of the study, as a PCA may severely underestimate the number of factors in the data if the goal is to determine the latent structure.

Limitations and Future Directions

While the current study provided an in-depth comparative analysis of the DETECT procedure, PCA, and common FA, it was not without its limitations. Most notably was the use of a single form of a math test. Each bootstrapped sample utilized in analysis represented a unique composite of the available students, however, both the assessment and the original examinee pool remained constant across the analyses. That is, the results of this study demonstrated how the dimensionality analyses differed across

sample sizes, but interpretation of those results remain within a relatively specific context. It would be beneficial for future research to repeat the study across multiple equivalent forms of an assessment, as well as across multiple assessments/assessment contexts.

Additionally, it was briefly mentioned in the summary of results that the DETECT procedure appeared to level out below the relative true number of factors, systematically under extracting. The current study did not result in any information that could provide an explanation for this finding. Future studies should consider fleshing out the potential under extraction problem in the DETECT procedure and consider evaluating the analysis from a finer grain size, such as the actual composition of the extracted factors rather than just the number of factors extracted.

References

Deng, N., Han, K. T., & Hambleton, R. K. (2013). A Review of DIMPACK Version 1.0: Conditional Covariance–Based Test Dimensionality Analysis Package. *Applied Psychological Measurement*, *37*(2), 162-172.

Dinno, A. (2018). paran: Horn's Test of Principal Components/Factors. R package version 1.5.2. https://CRAN.R-project.org/package=paran

Dinno, A. (2014). Gently clarifying the application of Horn's parallel analysis to principal component analysis versus factor analysis.

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The applications of exploratory factor analysis in applied psychology: Acritical review and analysis. *Personnel Psychology*, *39*, 291–314.

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal factors. *Journal of Vegetation Science*, *6*(1), 99-106.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, *7*(2), 191-205.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological measurement*, *66*(3), 393-416.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185.

Jang, E. E., & Roussos, L. (2007). An Investigation into the Dimensionality of TOEFL Using Conditional Covariance-Based Nonparametric Approach. *Journal of Educational Measurement*, *44*(1), 1-21.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, *20*(1), 141-151.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data (Doctoral dissertation, University of Illinois at Urbana-Champaign). Dissertation Abstracts International: Section, B, 5598, 12-55.

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best practices in exploratory factor analysis. *Best practices in quantitative methods*, 86-99.

Rencher, A. C. Methods of multivariate analysis. 2002. *DOI*, *10*(0471271357), 66.

Robitzsch A (2019). *sirt: Supplementary Item Response Theory Models*. R package version 3.7 40, https://CRAN.R-project.org/package=sirt.

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, *43*(3), 215-243.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical factor analysis to detect multidimensionality. *Journal of educational measurement*, *35*(1), 1-30.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research*, *99*(6), 323-338.

Singh, T., & West, N. E. (1971). Comparison of some multivariate analyses of perennial Atriplex vegetation in southeastern Utah. *Vegetatio*, *23*(5-6), 289-313.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589-617.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*(4), 331-354.

Svetina, D., & Levy, R. (2012). An overview of software for conducting dimensionality assessment in multidimensional models. *Applied Psychological Measurement*, *36*(8), 659-669

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321-327.

Watkins, M. W. (2006). Determining parallel analysis criteria. *Journal of modern applied statistical methods*, *5*(2), 344-346.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, *8*(3).

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika, 64, 213-249.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, *99*(3), 432.

Tables

Table 4.1

Measures of Accuracy and Precision by Method and Sample Size

	DETECT				PCA				Common FA				Mean	
N	Mode	ME	MAD	SD	Mode	ME	MAD	SD	Mode	ME	MAD	SD	MAD	SD
250	7	-7.420	7.420	1.486	2	-2.800	2.800	0.404	3	-11.640	11.640	0.964	7.287	0.951
500	10	-5.780	5.780	1.799	3	-2.400	2.400	0.606	5	-9.940	9.940	1.391	6.040	1.265
2500	12	-3.660	3.780	2.086	4	-0.700	0.700	0.463	11	-4.300	4.300	1.753	2.927	1.434
10000	12	-3.760	4.000	2.378	5	0.040	0.040	0.198	15	-0.160	0.760	1.095	1.600	1.224
25000	11	-3.660	3.740	2.246	5	0.220	0.220	0.418	18	2.320	2.320	0.935	2.093	1.200
Mean			4.944	1.999			1.232	0.418			5.792	1.228		

Figures

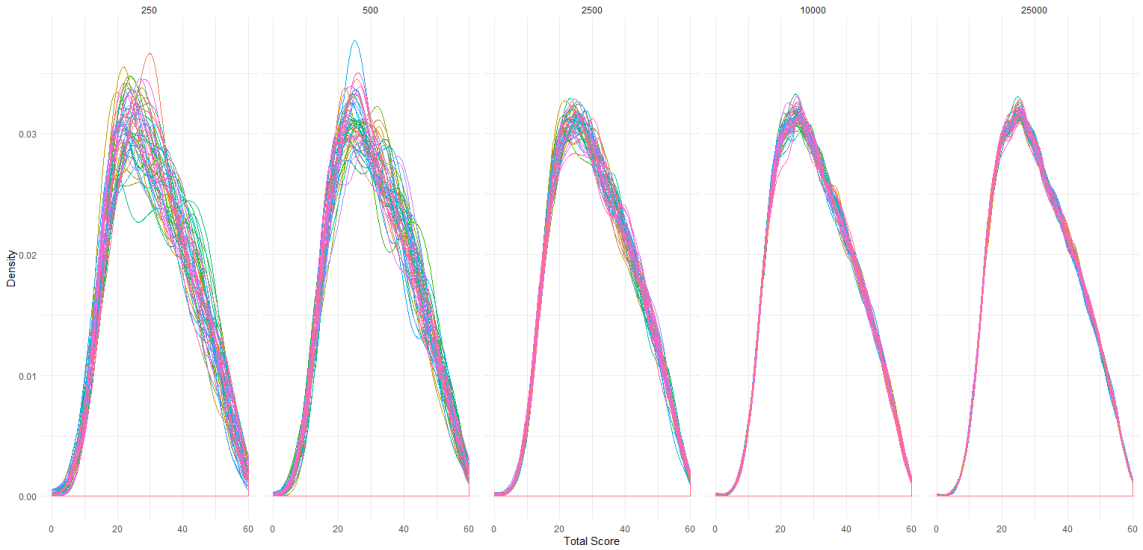

Figure 4.1. Overlaid Density Distributions of Total Score by Sample Size.

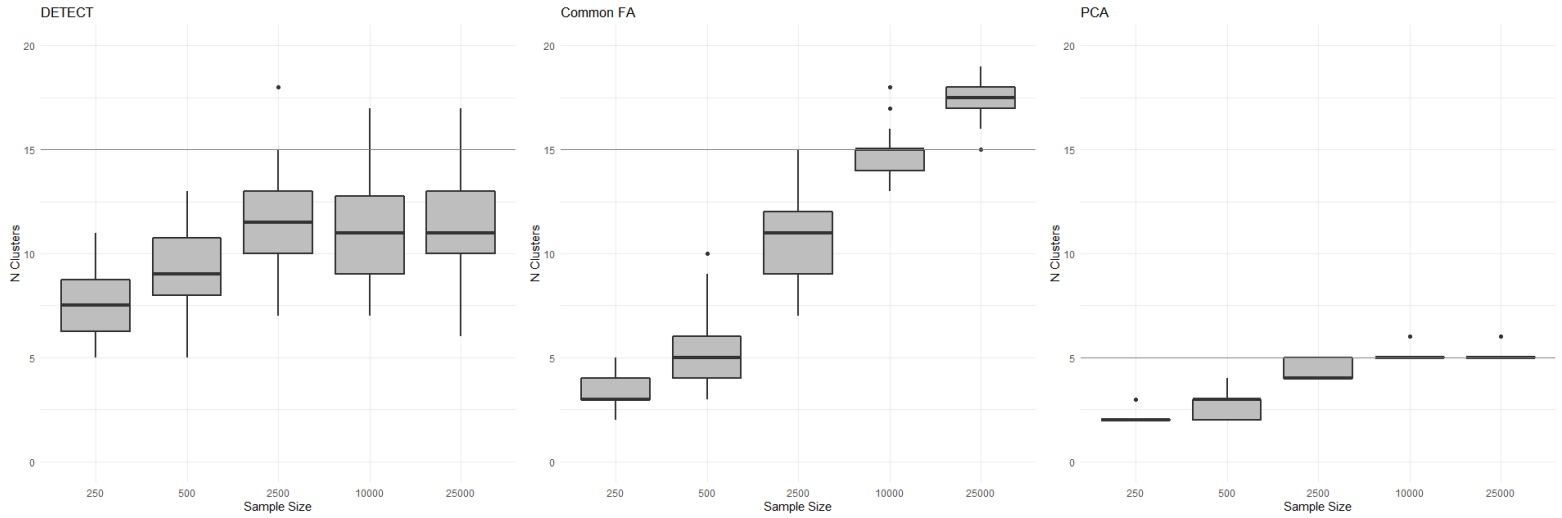

Figure 4.2. Factor Extraction Results by Method and Sample Size.

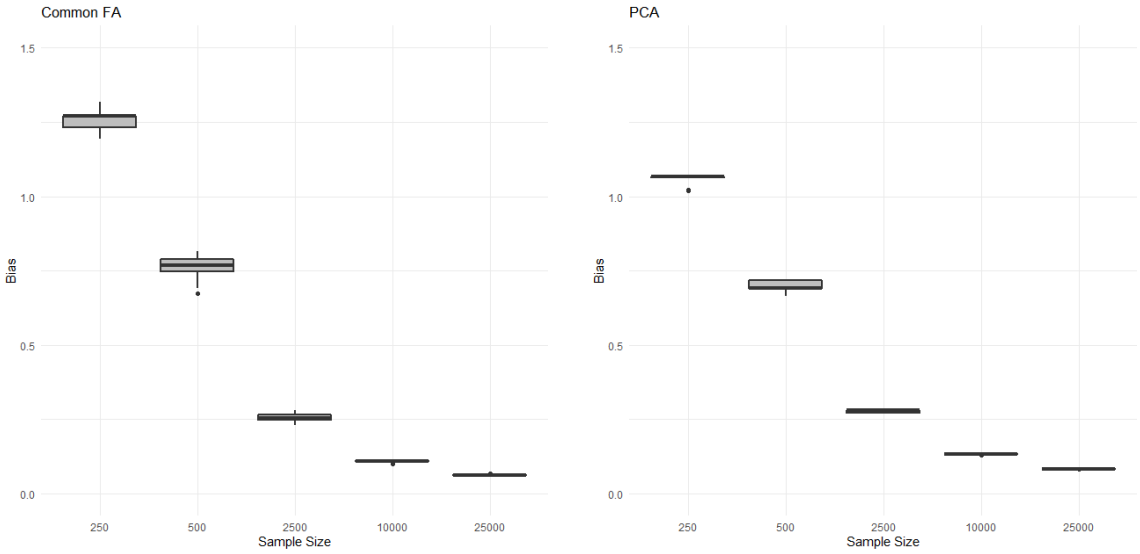

Figure 4.3. Bias Results by Method and Sample Size.

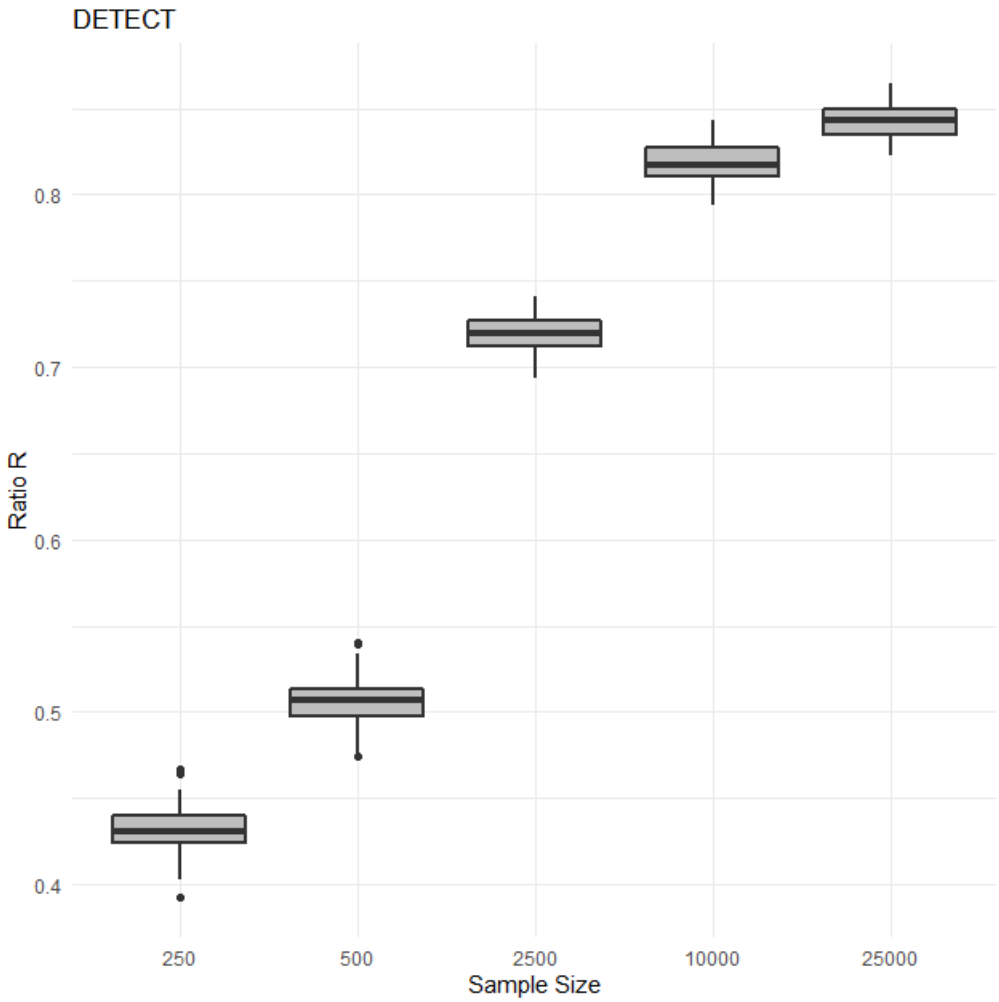

Figure 4.4. Ratio R Results by Sample Size.

CHAPTER V

PAPER 2: EVALUATING THE STABILITY OF FACTOR DEFINITIONS IN THE EXPLORATORY DETECT PROCEDURE

Introduction

The degree of stability in dimensionality analyses can have major implications when it comes to determining the validity of dimensionality results. Whether confirming or exploring dimensional structure in the data, the stability of those results directly impacts the degree of confidence we can have in using them in other decisions or forming interpretations. Moreover, the *Standards for Educational and Psychological Testing* state that when test scores and their interpretation depend on the "relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided" (American Educational Research Association [AERA], 2014, p. 27).

Stability may be defined in a variety of ways. From a confirmatory perspective, the stability of dimensionality results may refer to the consistent confirmation of previously defined dimensions across multiple administrations of an instrument. Additionally, stability may be assessed under a variety of conditions. For instance, some confirmatory analyses are employed across equivalent groups while others test across purposely-dissimilar groups. For example, Holzinger and Swineford (1939), tested the invariance of factors across administrations of various measures to two fundamentally different samples of participants where one sample was comprised of foreign-born respondents and one sample of native-born respondents. While invariance studies

encompass a wider range of sources of variance in results, the stability of methods used is a critical component of these analyses. Additionally, while some stability analyses utilize a single form of a specific instrument, others employ various versions of that instrument as Werneke, Goldberg, Yalcin, and Üstün (2000) demonstrated in assessing the factor structure of the General Health Questionnaire.

From an exploratory perspective, stability may refer to the consistent reproduction of an exploratory factor solution. That is, under various conditions, the ability to reproduce the same dimensionality results across those conditions. Because dimensionality is considered a function of test and examinee characteristics (Reckase, 1990), the strength of exploratory conclusions relies on their stability across conditions or administrations that are diverse in these characteristics. More specifically, in order to develop statistically defensible conclusions regarding the factor structure of a given test form, results of the exploratory analysis should be compared across examinee characteristics such as gender, grade, and language status (e.g. English language learners vs. native English speakers), and across test characteristics such as how highly correlated the underlying constructs being measured are, or characteristics specific to the testing context such as the number of examinees in the sample or the time of year the test was administered.

Previous research has documented the effects of various examinee characteristics on statistical test dimensionality. These investigations are highly concentrated in the areas of language assessment and testing. During evaluation of the Michigan English Language Assessment Battery (MELAB), Jiao (2004) found that results of test dimensionality

analyses varied by examinee gender, native language, and proficiency level. Investigations into the dimensionality of the widely utilized Test of English as a Foreign Language (TOEFL) reported similar findings. Mainly, the research suggests that dimensionality of the TOEFL is influenced by examinees' English language proficiency (Oltman, Stricker, & Burrows, 1988; Schedl, Gordan, Carey, & Tang, 1995). Other studies on the dimensionality of the TOEFL focused additionally on the influence of language background, ethnicity, and gender (Jang & Roussos, 2007). Previous research has demonstrated that the traits (e.g. gender, ethnicity, native language) and/or skills (e.g. proficiency) examinees possess and inherently contribute to the assessment situation may influence dimensionality and warrant greater investigation.

The effects of context-specific testing characteristics on dimensionality results are of primary interest in this study. By far, the most frequently researched of these characteristics is sample size, or the number of examinees available in the response data. In a study comparing the stability of three exploratory methods, maximum likelihood factor analysis, principal factor analysis, and rescaled image analysis, Velicer (1974) concluded that no one method clearly emerged as the most stable across an analysis of four different measures. Moreover, in a study by Guadagnoli and Velicer (1988), factor saturation, absolute sample size, and the number of items per factor were identified as important components of factor solution stability. Additionally, in a simulation study regarding the performance of the DETECT procedure, Tan and Gierl (2006, p.20-21) concluded that the "degree of complexity, correlation between dimensions, and item discrimination" affect the reliability of DETECT results.

While the factor structure of examinee response data is most commonly evaluated by employing common factor analysis or principle components analysis, other methods exist for assessing dimensionality. Exploratory DETECT is a sound alternative method for assessing dimensionality (Kim, 1994; Zhang & Stout, 1999). DETECT is a nonparametric conditional covariance-based procedure that assumes approximate simple structure in the data, and therefore attempts to map each item to the dimension (or factor) it predominately measures (Deng, Han, & Hambleton, 2012). When determining dimensional structure, "DETECT searches for a partitioning of the items into factors such that the conditional covariances between items from the same factor are positive and the conditional covariances between items from different factors are negative" (Jang & Roussos, 2007, pp. 6-7). Once DETECT reaches the optimum factor partition based on within and between factor conditional covariances, it outputs the resulting number or factors with accompanying item-to-factor mapping, and a variety of statistics that evaluate the appropriateness of the given dimensional output.

Although a less commonly used method, DETECT is widely advocated for and supported in the literature. DETECT may be preferable to parametric exploratory factor analytic methods because its nonparametric approach makes it comparatively less computationally intense, requiring only simplistic statistical computations instead of complex algorithms (Jang & Roussos, 2007; Roussos, Stout, & Marden, 1998; Roussos & Ozbek, 2006). Moreover, Roussos and Ozbek (2006) assert that "nonparametric techniques have become increasingly popular because they avoid strong parametric modeling assumptions while still adhering to the fundamental principles of item response

theory (IRT)" (p. 214). However, possibly the most practical advantage to using DETECT relates to its function of modeling approximate simple structure. That is, if approximate simple structure holds, sorting items into unique factors can aid in substantive interpretation of those factors, or as Zhang and Stout (1996) claim, the resulting factor partitions are "referred to as the *correct* [or] *dimensionally-based*" and they are considered "substantively meaningful and dimensionally separate" (Zhang & Stout, 1996, pp. 214; emphasis original).

Purpose

When addressing the stability of factor solutions, the current body of literature on statistical test dimensionality almost exclusively refers to the results provided by more traditional factor analytic approaches (e.g. EFA, PCA). The purpose of this study was to provide a fine-grained look at the stability of the nonparametric, conditional covariance-based procedure, exploratory DETECT and contextualize those results with a larger discussion regarding the effect of sample size and granularity of the focus of analyses.

Research Questions

1. How does sample size affect the stability of factor definitions of the exploratory DETECT procedure?
2. How does the effect of sample size on factor definition stability compare to the results regarding stability and accuracy of the number of factors determined by the exploratory DETECT procedure in previous research?

3. Is there an advantage to evaluating performance of the exploratory DETECT procedure at a smaller grainsize (i.e. factor composition level) or is the more parsimonious approach to evaluating the number of factors sufficient?

Methods

Data Sources

One form of a large-scale standardized math test was used in this study. The test contains 60 items spanning eight content areas and three cognitive skill domains. Response data is available for 25,000 examinees, but additional samples of varying size was composed for analysis.

Analyses

Prior to analysis, bootstrap resampling (i.e. sampling *with* replacement) was employed to construct the samples to be used in evaluating stability of the procedure. As the current study will serve as an extension of previous research by the author, sample sizes remained consistent between the studies. More specifically, five sets of samples were created to represent very small (*n=250*), small (*n= 500*), intermediate (*n=2,5000*), moderate (*n=10,000*) and large (*n=25,000*) sample sizes, each set containing 50 samples of the same size.

In addition to sample creation, baseline figures for "true" factor definitions were established. For the purpose of this study, the "true" factor definitions will refer to the item-to-factor mapping provided by the exploratory DETECT procedure when applied to the original sample of response data from the 25,000 examinees. This mapping served as the key for determining stability across the various sets of constructed samples.

As noted, dimensionality was evaluated using the nonparametric, conditional covariance-based statistical procedure, exploratory DETECT (Kim, 1994; Zhang & Stout, 1999). DETECT provides a host of statistics to determine the magnitude of multidimensionality in the data, with the primary statistic being the DETECT value. The DETECT value reports the mean of all item-pair conditional covariances as a measure of the magnitude of multidimensionality (Roussos & Ozbek, 2006), and is calculated as follows:

$$D(P, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \le i \le ln} \delta_{i,l} E[cov(U_i, U_l | \Theta_{TT})] \quad (1)$$

where $\delta_{i,l} = 1$ if item i and l are in the same factor and -1 if they are in different factors, *P* represents the given partition of items in the assessment, *n* is the number of items on the assessment, *i and l* are any two items on the assessment with scores U_i and U_l, and Θ_{TT} represents "the unidimensional latent variable 'best measured' by the total test [number correct] score" (Stout et al., 1996, p. 333). DETECT index values less than or equal to .1 indicate probable unidimensionality while values greater than 1.0 suggest the presence of multidimensionality (Stout, 1987; Stout et al., 1996). For the DETECT index estimate to reach its maximum possible value, all within-factor item pair covariances must be positive while all between factor item pair covariances are negative (Stout et al., 1996). The maximum DETECT value may be obtained by:

$$D^*(P, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \le i \le ln} |E[cov(U_i, U_l | \Theta_{TT})]|, \quad (2)$$

Additionally, DETECT provides the IDN index and ratio r. The ratio r compares the computed DETECT value to its maximum possible value. The ratio is computed by

$$r = \frac{D(P,\Theta_{TT})}{D^*(P,\Theta_{TT})} \quad (3)$$

The higher the ratio the more confidently multidimensionality can be concluded. The maximum value of 1 indicates simple structure, while .8 suggests approximate simple structure (Svetina & Levy, 2012). Finally, the IDN index can be defined as the proportion of within-factor conditional covariance values that were positive. Again, the higher the value the more confidently we can conclude multidimensionality. DETECT was used to assess the dimensionality of all constructed samples and employed the *sirt* (Robitzsch, 2019) package in the R programming environment.

Once exploratory DETECT analyses were completed, stability was evaluated at each sample size by calculating the average deviation from the "true" factor-mapping, with lower average deviations representing higher stability. That is, for results of each iteration of the exploratory DETECT analysis, factors were matched with the "true" solution aiming to maximize the number of matched items within a factor. All factors from the sample solution were compared with each true factor. The matching process moved from the largest factor to the smallest factor for the total *n* factors of each iteration, prioritizing maximum matching for factors of larger size, however, if a smaller factor existed with a greater than or equal to number of matches for a specified true factor, that smaller sample factor won the match.

Once factor matches were established, item-factor mapping was analyzed. For each item-factor match the value of item deviation equals 0, while each item from the sample solution that goes unmatched received a deviation value of 1. More specifically, if an item appeared in a different factor in the sample solution than the true solution, that item only received a single item deviation value of 1, rather than counting as a mismatch in both the sample and true solutions. Moreover, if the number of factors in the sample solution exceeded the number of true factors, all items in the remaining unmatched factors received an item deviation value of 1. Therefore, total DETECT factor congruence (DFC) is conceptualized as one minus the sum of item deviations divided by the total number of items in the test in a given exploratory DETECT factor solution:

$$DFC = 1 - \frac{\sum_{i=1}^{n} ID_i}{n}, \quad (4)$$

where *i* is a given item on the test, *n* is the total number of items on the test, and *ID* is the item deviation value of 0 or 1. Additionally, to further assess stability, DETECT statistics of each sample are reported. While DFC values were used to inform interpretations of how sample size impacts the stability of DETECT factor definitions, the DETECT statistics provided additional information on the relationship between DFC and the degree of multidimensionality in the data.

Results

Data

As previously noted, the original dataset consisted of responses from 25,000 examinees on 60 items. The mean total score from the original dataset was 30.87

(*sd=11.65*), indicating that the average student answered just over 50% of the questions correct. Additionally, the median total score was 30, slightly lower than the mean, and visual inspection of the density distribution of total scores indicated that the data is very slightly positively skewed. Consistency in sampling distributions was also inspected, with samples increasing in consistency as sample size increases, but remaining within reasonable deviation at the lower sample sizes as well. Figure 5.1 provides a visual of the sample consistency results.

Defining Truth

An initial analysis was completed employing the DETECT procedure with the original 25,000 observations. The goal of this analysis was to establish a relative "true" item-to-factor mapping. The exploratory DETECT procedure requires a user-defined maximum number of factors to estimate. This user-defined value can range from two to the total number of items in the data. For the initial analysis the max number of factors to be estimated was freed to equal 60, the total number of items in the data. Initial results of the exploratory DETECT procedure indicated there were 15 factors in the data. After establishing truth, the analysis was run on all 50 samples of each size. All subsequent DETECT analyses employed a maximum number of factors to estimate of 20 based on the baseline results of 15.

Factor Size and Difficulty

Of the resulting 15 factors from the baseline analysis, the minimum factor size was 2 items and the maximum factor size was 7 (Table 5.1). Most factors (50%) consisted of just three items. Factor difficultly, defined in this study as the mean

proportion correct across items in the factor, ranged from .24 to .84 with a mean of .50 and a median value of .48 (Table 5.1).

DETECT Factor Congruence

As sample size increased, DFC values also increased resulting in a maximum mean DFC at the original sample size of 25,000. The minimum average DFC value was 0.332 for samples of size 250, indicating that, on average, there was about 33% agreement between the baseline and sample item-to-factor mappings. The maximum average DFC value was 0.668 for samples of size 25,000, which could again be interpreted as an average of about 67% agreement between the baseline and sample item-to-factor mappings. In terms of variability in DFC values within samples sizes, samples of size 250 showed the least variability in results with a standard deviation across samples of .044 and samples of size 2,500 showed the greatest variability with a standard deviation of .072. See Figure 5.2 for full DFC results by sample size.

The minimum DFC value across all samples was .217, occurring at the sample size of 250, indicated that the minimum overlap between the baseline results and the item-to-factor mapping from all 250 samples of various sizes was approximately 22%. The maximum DFC value across all samples was 0.817, occurring at the sample size of 25,000, indicating that the maximum overlap from all 250 samples was approximately 82%. Moreover, the maximum FC value indicated that even when sample sizes were consistent with the original data, the exploratory DETECT procedure failed to perfectly replicate the baseline results in any of the 50 bootstrapped samples of that size.

Results by Grain Size

As noted throughout, a previous study was conducted evaluating the accuracy and precision of the number of factors extracted by the DETECT procedure at various sample sizes (Lay, 2020a) The previous study utilized the same sample sizes as the present study and employed the mean absolute difference and the mean error to evaluate the degree and direction of the accuracy of the number of factors extracted, and the standard deviations to evaluate precision. The results of the previous study are relevant to the present study as they allow for a comparative analysis of the performance of the exploratory DETECT procedure at different levels of granularity. While the results cannot be evaluated as direct comparisons, the measures of accuracy from the previous study can be reasonably compared to the FC values of the present study as the underlying concept of both measures is to evaluate the similarity of baseline and sample results. Likewise, the standard deviations from both studies can be compared to evaluate the consistency of sample results by sample size.

Across the first three sample sizes, the pattern of the resulting MAD and mean DFC values appeared to be consistent. That is, regardless of grain size, as sample size increased, similarity between the baseline results and the sample results increased. However, after samples of size 2,500, the pattern of results begins to differ between studies. More specifically, the MAD results from the previous study appear to level out, having extremely similar values across sample sizes of 2,500, 10,000, and 25,000 while the mean DFC values continue to increase as sample size increases. It is worth noting, however, that while DFC values continue to increase with sample size, the range of

differences between the means are considerably smaller (0.558-0.668) than the differences between the smaller sample sizes (0.332-0.558).

Regarding variability of results by sample sizes, both studies resulted in the least variability in results for samples of size 250. For the number of factors extracted at a sample size of 250, the corresponding standard deviation was 1.486, while the standard deviation of DFC values at the same sample size was 0.044. However, different sample sizes exhibited maximum variability for the two studies. That is, the previous study demonstrated the greatest variability for samples of size 10,000 with a standard deviation of 2.378, while the current study exhibited the greatest variability in results at the sample size of 2,500 with a standard deviation of 0.072. Table 5.2 provides the accuracy and precision results from the previous study and Figure 5.3 provides the DETECT factor extraction results.

DETECT Statistics

Results of the DETECT values by sample size indicated that as sample size increased, the DETECT value generally decreased, representing a decrease in the strength of multidimensionality in the data with an increase in the number of examinees included in analysis. However, the differences in mean DETECT values by sample size were relatively small, and with the exception of some of the results at a sample size of 250, there was no interpretive difference of results between sample sizes. More specifically, while some of the resulting DETECT values from the samples of size 250 could be interpreted as indicating moderate dimensionality by exceeding the threshold of .4, all other values fell between .2 and .4, representing a consistent interpretation of weak

dimensionality across sample sizes. Additionally, the DETECT values from the current study appear to follow the inverse of the pattern of factor extraction results from the previous study. That is, while the DETECT values notably decreased between the sample sizes of 250 and 2,500, they appear to level out from 2,500 to 25,000.

Results of the ratio R values by sample size provided a contrasting picture of the relationship between sample size and the degree of multidimensionality provided by the exploratory DETECT procedure. Unlike the DETECT values, the ratio R values generally increased as sample size increased rather than eventually leveling out, consistent with the previous study. However, as with the resulting DETECT values, there were no interpretative differences in the ratio R values across sample sizes. That is, all values exceeded the threshold of .36, indicating an essential deviation from unidimensionality. Moreover, while the DETECT values appeared to follow the inverse pattern of the results from the previous study, the ratio R values appeared to follow the pattern of the DFC results from the current study. Figure 5.4 provides the DETECT values and ratio R values by sample size.

The inconsistent relationships of the DETECT and ratio R values to sample size prompted a further investigation into the relationship between the two statistics. That is, one would reasonably expect that as the DETECT value increases the ratio R value would increase given that both are indicators of the degree of multidimensionality in the data, but when evaluating the statistics across sample sizes, one increased while the other decreased. To evaluate the relationship further, results were explored within sample sizes. Once the DETECT and ratio R values were plotted against each other within sample

sizes, results supported the expectation that as one increased the other would as well despite their differential relationships by sample size. See Figure 5.5 for the results of the DETECT values by ratio R values.

Discussion

Main Findings

Unsurprisingly, the baseline item-to-factor mapping was more stable in samples of similar size, as demonstrated by the DFC results. That is, as the sample size increased, DFC values also increased. This is likely related to the factor extraction results from the previous study. Demonstrated by the previous study, as sample size increased towards the baseline size of 25,000, the number of factors extracted generally increased up to the sample size of 2,500. This could partly explain the increase in factor congruence values as the closer the number of factors is to relative truth, the greater the opportunity for overlap in factors between relative truth and sample output. However, the number of factors extracted can only be partly responsible for the increase in factor congruence values by sample size as where the number of factors extracted reaches a leveling out point after a sample size of 2,500, the factor congruence values continue to increase across all 5 sample sizes.

As previously noted, the results of the DETECT values by sample size appeared to be the direct inverse of the factor extraction results from the previous study. The relationship between the two sets of results complicated the interpretation of the relationship between the DETECT values and sample size. That is, the previous study concluded that the number of factors extracted was, at least in part, related to sample size,

while the current study resulted in the same conclusion regarding the DETECT values and sample size. However, when evaluating the two sets of results together, it is unclear whether the DETECT values are more strongly related to sample size or the number of factors extracted.

Additionally, it was noted that the pattern of the factor congruence values by sample size and ratio R values by sample size were extremely similar. That is, both statistics continued to increase as sample size increased. These results add an interesting layer to the narrative regarding the relationship between dimensionality, as conceptualized in the DETECT procedure, and sample size. That is, although it appears that not much changes in terms of the number of factors extracted and the DETECT value after a sample size of 2,500, the degree of similarity between the estimated DETECT value and the maximum possible DETECT value continues to increase, resulting in a higher ratio R, while the DFC values continue to increase as well. This may warrant additional research into the performance of the maximum DETECT value by sample size, and its relationship to the number of factors extracted as well as FDC values, as it implies that the max DETECT is actually decreasing while the estimated values hold constant.

DETECT Performance by Grain Size

While both studies provided useful information regarding the relationship between sample size and dimensionality, the results of the current study suggest that going beyond the number of factors extracted, and instead evaluating the actual item-to-factor mapping, provides a greater degree of accuracy when evaluating sample results to relative truth. That is, the leveling out of the number of factors extracted and DETECT

values may camouflage the gained stability in the more fine-grained analysis of item-to-factor overlap in the higher sample sizes. However, it is worth noting that if the object of the study was to purely determine the number of factors in the data or evaluate the degree of dimensionality using the DETECT value, as are the most typical uses of the DETECT procedure, results of the two studies suggest that it may require less examinee data to reach stability and the more fine grained analysis could be avoided.

Limitations and Future Research

While the current study provided valuable information regarding the stability of the exploratory DETECT item-to-factor mapping across sample sizes, it also exposed areas that necessitate additional research. For one, despite factor congruence values steadily increasing with sample size, they never exceeded 0.817. Further research should be done with samples of a larger size to evaluate whether there is a point at which the values stabilize and/or average closer to 1. Additionally, due to the results of the DETECT and ratio R values, a further investigation into the behavior of the max DETECT value by sample size is necessitated.

Finally, considering both the current and previous study relied on the same base sample of examinee response data, generalizability of the results is limited. Future research should focus on expanding the current study to additional assessments that vary in terms of content, composition, and length. Moreover, it would be beneficial to replicate the analyses included in the current and previous studies across multiple forms of an assessment.

References

American Educational Research Association, American Psychological Association, Joint

Committee on Standards for Educational, Psychological Testing (US), & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Deng, N., Han, K. T., & Hambleton, R. K. (2013). A Review of DIMPACK Version 1.0: Conditional Covariance–Based Test Dimensionality Analysis Package. *Applied Psychological Measurement*, *37*(2), 162-172.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of factor patterns. *Psychological bulletin*, *103*(2), 265.

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi factor solution. *Supplementary Educational Monographs*.

Jang, E. E., & Roussos, L. (2007). An Investigation into the Dimensionality of TOEFL Using Conditional Covariance-Based Nonparametric Approach. *Journal of Educational Measurement*, *44*(1), 1-21.

Jiao, H. (2004). Evaluating the dimensionality of the Michigan English language assessment battery. *Spaan Fellow Working Papers in Second or Foreign Language Assessment Volume 2 2004*, *1001*, 27.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data (Doctoral dissertation, University of Illinois at Urbana-Champaign). Dissertation Abstracts International: Section, B, 5598, 12-55.

Lay, A. L. (2020a). *A Comparative Study of the Accuracy and Precision of Exploratory Dimensionality Analyses* (doctoral dissertation excerpt, The University of North Carolina at Greensboro).

Oltman, P. K., Stricker, L. J., & Barrows, T. S. (1988). Native language, English proficiency, and the structure of the Test of English as a Foreign Language. *ETS Research Report Series*, *1988*(1), i-36.

Reckase, M. D. (1990). Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests.

Riegle-Crumb, C. (2006). The path through math: Course sequences and academic performance at the intersection of race-ethnicity and gender. *American Journal of Education*, *113*(1), 101-122.

Robitzsch A (2019). *sirt: Supplementary Item Response Theory Models*. R package version 3.7 40, https://CRAN.R-project.org/package=sirt.

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, *43*(3), 215-243.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical factor analysis to detect multidimensionality. *Journal of educational measurement*, *35*(1), 1-30.

Schedl, M., Gordon, A., Carey, P. A., & Tang, K. L. (1995). An analysis of the dimensionality of TOEFL reading comprehension items. *ETS Research Report Series*, *1995*(2), i-26.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589-617.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*(4), 331-354.

Svetina, D., & Levy, R. (2012). An overview of software for conducting dimensionality assessment in multidimensional models. *Applied Psychological Measurement*, *36*(8), 659-669.

Tan, X., & Gierl, M. J. (2006, April). Evaluating the consistency of DETECT indices and item factors using simulated and real data that display both simple and complex structure. In *annual meeting of the American Educational Research Association, San Francisco, CA*.

Velicer, W. F. (1974). A comparison of the stability of factor analysis, principal factor analysis, and rescaled image analysis. *Educational and Psychological Measurement*, *34*(3), 563-572.

Watkins, M. W. (2006). Determining parallel analysis criteria. *Journal of modern applied statistical methods*, *5*(2), 344-346.

Werneke, U., Goldberg, D. P., Yalcin, I., & Üstün, B. T. (2000). The stability of the factor structure of the General Health Questionnaire. *Psychological Medicine*, *30*(4), 823-829.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3).

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213-249.

Tables

Table 5.1

Factor Size and Difficulty

Factor	N Items	Difficulty
1	7	.68
2	3	.84
3	7	.60
4	3	.72
5	4	.65
6	3	.65
7	6	.57
8	3	.48
9	3	.41
10	4	.31
11	5	.35
12	2	.41
13	3	.26
14	3	.33
15	4	.24

Table 5.2

Measures of Accuracy and Precision by Method and Sample Size

	DETECT				PCA				Common FA				Mean	
N	Mode	ME	MAD	SD	Mode	ME	MAD	SD	Mode	ME	MAD	SD	MAD	SD
250	7	-7.420	7.420	1.486	2	-2.800	2.800	0.404	3	-11.640	11.640	0.964	7.287	0.951
500	10	-5.780	5.780	1.799	3	-2.400	2.400	0.606	5	-9.940	9.940	1.391	6.040	1.265
2500	12	-3.660	3.780	2.086	4	-0.700	0.700	0.463	11	-4.300	4.300	1.753	2.927	1.434
10000	12	-3.760	4.000	2.378	5	0.040	0.040	0.198	15	-0.160	0.760	1.095	1.600	1.224
25000	11	-3.660	3.740	2.246	5	0.220	0.220	0.418	18	2.320	2.320	0.935	2.093	1.200
Mean			4.944	1.999			1.232	0.418			5.792	1.228		

Figures

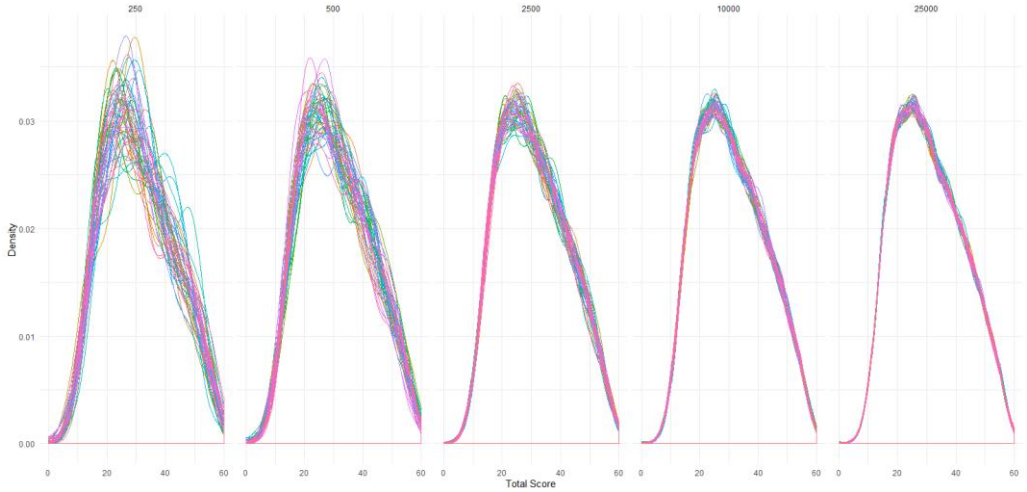

Figure 5.1. Overlaid Density Distributions of Total Score by Sample Size.

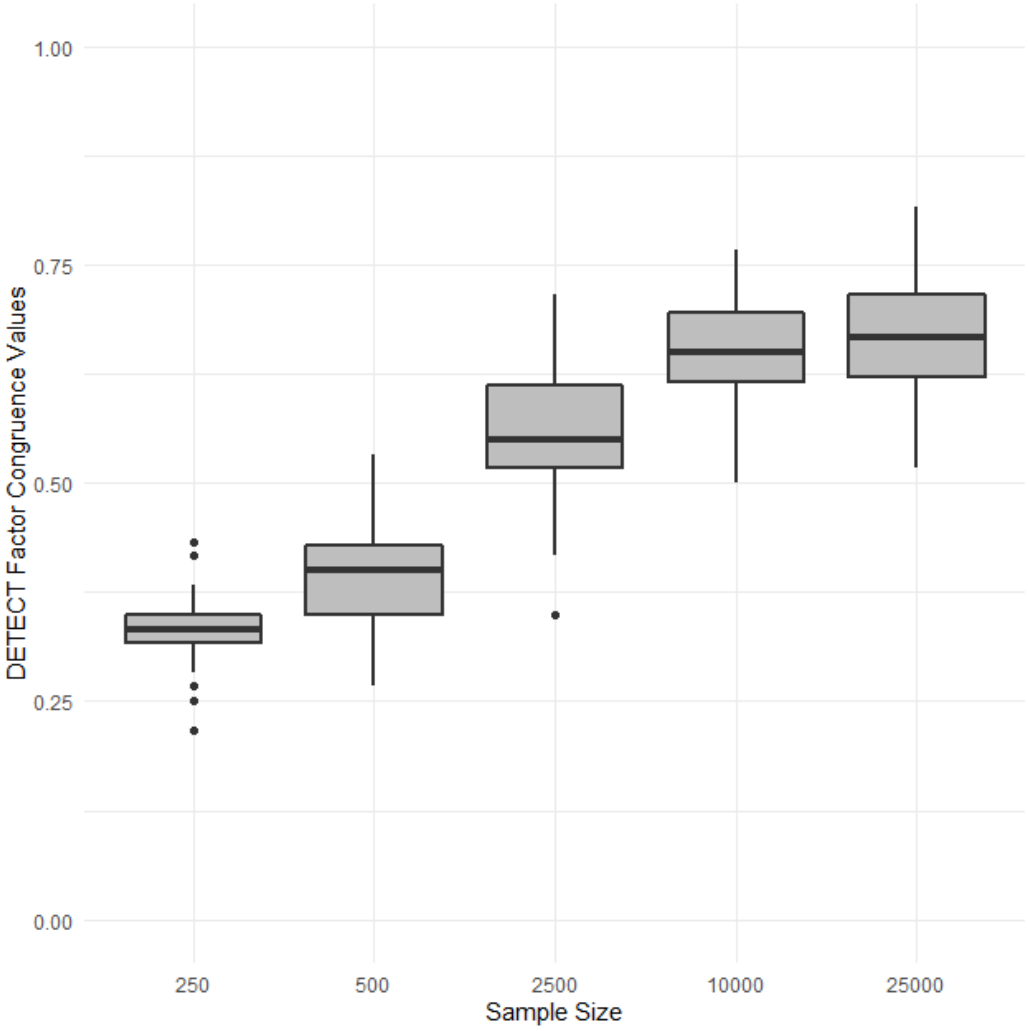

Figure 5.2. DETECT Factor Congruence Values by Sample Size.

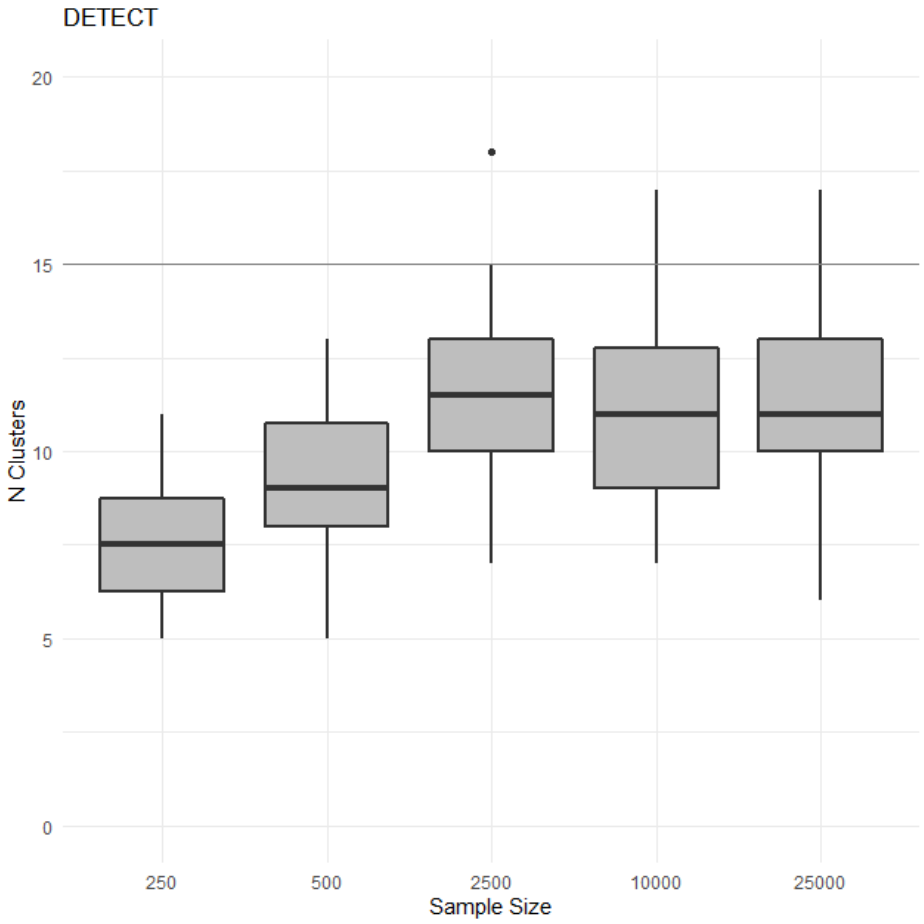

Figure 5.3. DETECT Factor Extraction Results by Sample Size.

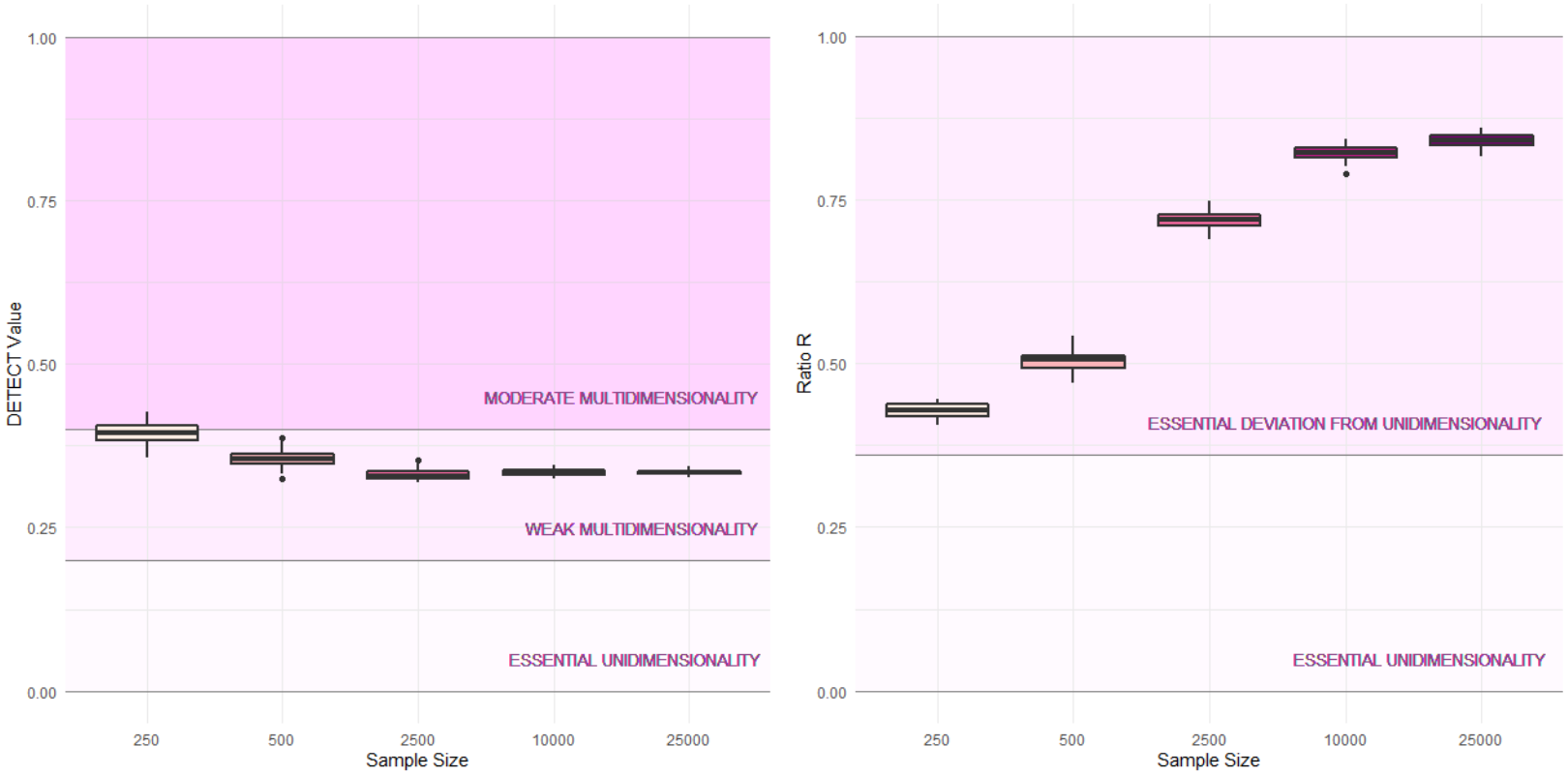

Figure 5.4. DETECT Values and Ratio R Values by Sample Size.

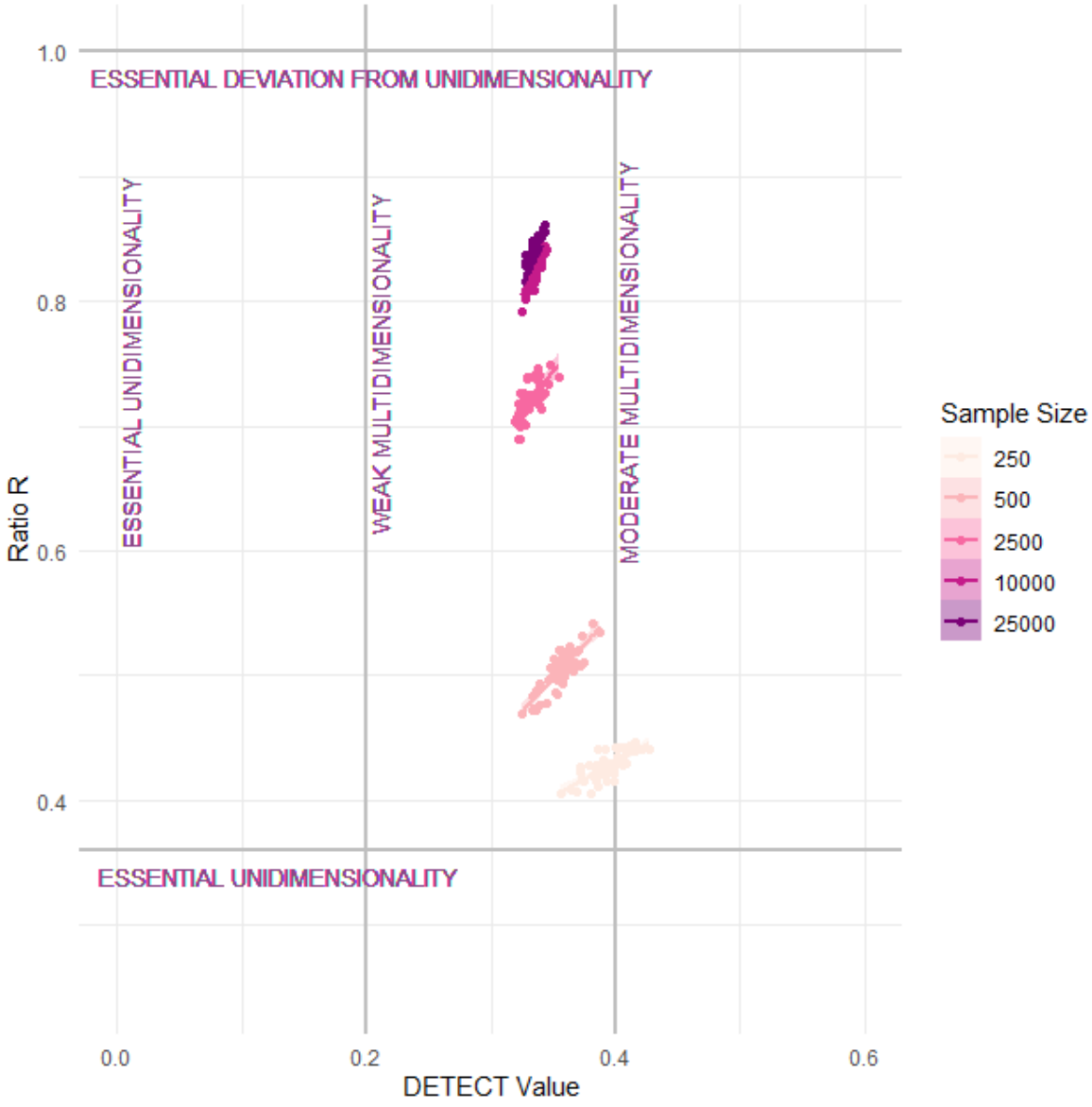

Figure 5.5. DETECT Values by Ratio R Values.

CHAPTER VI

PAPER 3: BUILDING DISCRIMINANT VALIDITY EVIDENCE FOR EXPLORATORY FACTORS WITH CONFIRMATORY ANALYSES

Introduction

With the consistent emergence of ever-evolving, complex assessments, the consideration of test dimensionality is rapidly regaining attention from the field of psychometrics. While some researchers aim to explore the factor structure of their assessment, others attempt to confirm the delineation of a factor structure developed *a priori*. Unfortunately, little research exists for evaluating dimensionality from both exploratory and confirmatory perspectives in tandem as a technique for strengthening validity evidence of the underlying factor structure.

For those focused on dimensionality exploration, the resulting factor structure necessitates a thorough validity investigation prior to operational and/or research related use and interpretation of factor-specific sub scores. One approach to evaluating the validity of different factors within a single assessment is to establish discriminant validity evidence between factors. Discriminant validity evidence refers to statistical results supporting the dissimilarity of two or more factors and "is particularly critical for discounting plausible rival alternatives to the focal construct interpretation" (Messick, 1995, p. 746). Moreover, the *Standards for Educational and Psychological Testing* state that when test scores and their interpretation depend on the internal structure of the assessment, or the "relationships among test items or among parts of the test, evidence

concerning the internal structure of the test should be provided," citing factor analysis as a potential method for establishing such evidence (AERA, 2014, p. 27).

A variety of methods exist for evaluating statistical test dimensionality from a confirmatory perspective. Of those methods, the most commonly cited technique is confirmatory factor analysis (CFA). Unlike exploratory methods, which are defined by the data, CFA requires a pre-defined factor structure and specified correlations between factors prior to analysis. Once the model is specified according to the hypothesized factor structure, the analysis is run to determine the degree of statistical confidence that may be concluded of the specified structure.

The use of CFAs is widely documented in the literature. Most commonly, CFAs are used to test the defensibility of theoretical constructs underpinning the measure at hand or the statistical differentiation of domains in attitudinal assessments. McAuley and Duncan (1989) tested the five-factor structure hypothesized to be underlying the Intrinsic Motivation Inventory. Likewise, McCrae, Zonderman, Costa Jr, Bond, and Paunonen (1996) evaluated the factor structure of the NEO Personality Inventory. Additionally, Mueller, Lambert, and Burlingame, (1998) evaluated the factor structure by content domains in the Outcome Questionnaire while Bosscher and Smit (1998) investigated the three factors alleged to makeup the General Self-Efficacy Inventory.

In addition to traditional factor analytic techniques, an alternative method for determining the dimensionality of examinee response data is DIMTEST (Stout, 1987). DIMTEST is a nonparametric, conditional covariance-based procedure for assessing dimensionality. DIMTEST is considered a test for unidimensionality (Nandakumar & Yu,

1996). That is, DIMTEST tests for unidimensionality between two user-defined subsets of data, referred to as the "assessment subset" (AT) and "partitioning subset" (PT) (Finch & Habing, 2007). If the resulting DIMTEST statistical has a p-value less than or equal to .05, the subsets are considered dimensionally distinct from one another and the presence of multidimensionality may be claimed.

Support for the use of DIMTEST is provided by previous research. Of the two "covariance-based, non-factor-analytic procedures" reviewed by Seraphine (2000), the researcher notes it as "the more promising" of the two (p. 83). Moreover, studies have demonstrated that DIMTEST remains sensitive to detecting dimensional distinctness while still controlling the Type 1 error rate (Nandakumar, 1994; Seraphine et al., 1995). Finally, Nandakumar and Yu (1996) demonstrated the consistency of the DIMTEST statistic across six different ability distributions, a normal distribution and five non-normal distributions through the use of a simulation study. While DIMTEST can be used in exploratory or confirmatory mode, this study focuses solely on utilizing DIMTEST for confirmatory purposes.

Another alternative conceptualization of statistical dimensionality is to consider it from an exploratory perspective rather than a confirmatory one. One available method for assessing dimensionality from this alternative perspective is the exploratory DETECT procedure (Kim, 1994; Zhang & Stout, 1999). DETECT, like DIMTEST, assumes approximate simple structure in the data, and therefore attempts to map each item to the dimension (or factor) it predominately measures (Deng, Han, & Hambleton, 2012). When determining dimensional structure, "DETECT searches for a partitioning of the items into

factors such that the conditional covariances between items from the same factor are positive and the conditional covariances between items from different factors are negative" (Jang & Roussos, 2007, pp. 6-7). Once DETECT reaches the optimum factor partition based on within and between factor conditional covariances, it outputs the resulting number or factors with accompanying item-to-factor mapping, and a variety of statistics that evaluate the appropriateness of the given dimensional output.

The use of DETECT is widely advocated for and supported in the literature. One reason DETECT is advocated for over parametric exploratory factor analytic methods is that being nonparametric makes it comparatively less computationally intense, requiring only simplistic statistical computations instead of complex algorithms (Jang & Roussos, 2007; Roussos, Stout, & Marden, 1998, Roussos & Ozbek, 2006). Moreover, Roussos and Ozbek (2006) assert that "nonparametric techniques have become increasingly popular because they avoid strong parametric modeling assumptions while still adhering to the fundamental principles of item response theory (IRT)" (p. 214). However, possibly the most practical advantage to using DETECT relates to its function of modeling approximate simple structure. That is, if approximate simple structure holds, sorting items into unique factors can aid in substantive interpretation of those factors, or as Zhang and Stout (1996) claim, the resulting factor partitions are "referred to as the correct [or] dimensionally-based" and they are considered "substantively meaningful and dimensionally separate" (Zhang & Stout, 1996, pp. 214; emphasis original).

While the exploratory DETECT procedure is extremely useful in investigating the dimensional structure of an assessment, alone it cannot provide sufficient validity

evidence for the resulting factor structure. Instead, combining the aforementioned methods to evaluating dimensionality, where each method represents a single step in a sequence of analysis, could prove useful for building discriminant validity evidence. More specifically, when utilizing the exploratory DETECT procedure to determine a statistically distinct set of factors within an assessment, confirmatory DIMTEST could subsequently be applied to those factors as a post hoc method for evaluating the independence of those factors from each other, or in other words, an attempt to build discriminant validity evidence.

Purpose

While the current body of literature on evaluating statistical test dimensionality within the context of educational assessment continues to become more saturated with overviews of available methods and applicative examples of those methods in use, most of this research fails to extend the discussion to validity considerations. The purpose of this study was to center the process of evaluating statistical dimensionality around building validity evidence by providing explicative examples of how multiple confirmatory dimensionality analyses can be used in building discriminant validity evidence for a predefined, exploratory factor structure.

Research Questions

1. Do confirmatory results from CFA and DIMTEST tend to support the factor structure provided by exploratory DETECT?
2. Did one confirmatory technique recover the DETECT factor structure better than the other?

3. What is the effect of sample size on the effectiveness of using these techniques to build discriminant validity evidence for the DETECT-specified factor structure?

Methods

Data Sources

One form of a large-scale standardized math test was used in this study. The test contains 60 items spanning eight content areas and three cognitive skill domains. Response data is available for 25,000 examinees, but additional samples of varying size were composed for analysis. More specifically, following the procedure for sample construction laid out in previous relevant research (Lay, 2020a; Lay, 2020b), five sets of samples were created to represent very small (n=250), small (n= 500), intermediate (n= 2,500), moderate ($n=10,000$) and large ($n=25,000$) sample sizes, using bootstrap sampling (i.e. sampling *with* replacement), each set containing 50 samples of the same size.

Analyses

This study was completed as a continuation of a previous set of exploratory dimensionality studies (Lay, 2020a; Lay, 2020b). That is, previously a variety of exploratory methods were used to determine the factor structure of the data at hand. While the goal of those studies was to gain a better understanding of the accuracy and stability of exploratory methods of determining the factor structure of examinee response data, with an emphasis on the effect of sample size, this study shifted to a confirmatory perspective. More specifically, this study employed confirmatory DIMTEST (Stout,

Douglas, Junker, & Roussos, 1993) to be used with the previously defined exploratory DETECT factors.

For additional context, a brief technical overview of the exploratory DETECT procedure (Kim, 1994; Zhang & Stout, 1999) is provided to understand how the factor structure was derived in previous research. When employed to assess dimensionality of examinee response data, the DETECT procedure provides a host of statistics to determine the magnitude of multidimensionality in the data. The primary statistic of interest reported by the analysis is the DETECT value. The DETECT value reports the mean of all item-pair conditional covariances as a measure of the magnitude of multidimensionality (Roussos & Ozbek, 2006), and is calculated as follows:

$$D(P, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \le i \le ln} \delta_{i,l} E[cov(U_i, U_l | \Theta_{TT})] \quad (1)$$

where $\delta_{i,l} = 1$ if item i and l are in the same factor and -1 if they are in different factors, *P* represents the given partition of items in the assessment, *n* is the number of items on the assessment, i and l are any two items on the assessment with scores U_i and U_l, and Θ_{TT} represents "the unidimensional latent variable 'best measured' by the total test [number correct] score" (Stout et al., 1996, p. 333).

DETECT index values less than or equal to .1 indicate probable unidimensionality while values greater than 1.0 suggest the presence of multidimensionality (Stout, 1987; Stout et al., 1996). For the DETECT index estimate to reach its maximum possible value, all within-factor item pair covariances must be positive while all between factor item pair

covariances are negative (Stout et al., 1996). The maximum DETECT value may be obtained by:

$$D^*(P, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \le i \le ln} |E[cov(U_i, U_l | \Theta_{TT})]|. \quad (2)$$

Additionally, DETECT provides the IDN index and ratio r. The ratio r compares the computed DETECT value to its maximum possible value. The ratio is computed by

$$r = \frac{D(P, \Theta_{TT})}{D^*(P, \Theta_{TT})}. \quad (3)$$

The higher the ratio the more confidently multidimensionality can be concluded. The maximum value of 1 indicates simple structure, while .8 suggests approximate simple structure (Svetina & Levy, 2012). Finally, the IDN index can be defined as the proportion of within-factor conditional covariance values that were positive. Again, the higher the value the more confidently we can conclude multidimensionality.

As previously noted, both exploratory DETECT and confirmatory DIMTEST are nonparametric, conditional covariance-based dimensionality procedures that assumes approximate simple structure in the data. Because of this, it is a natural progression to plug the DETECT factor structure results into the DIMTEST analysis. That is, the actual item-to-factor mapping obtained via exploratory DETECT was be used to define distinct factor pairings to be tested against each other via confirmatory DIMTEST. In the confirmatory DIMTEST procedure, two subsets of data are tested against each other at a time for dimensional distinctness. Those two subsets are conceptualized in DIMTEST as the assessment subtest (AT) and the partitioning subtest (PT) and were represented by the

DETECT factor pairings in this study. In the previous research, the factor definitions from the full sample (*n=25,000*) were used to define "truth" and those definitions remain consistent for this study. Factors were tested against each other for dimensional distinctness in every iteration of the constructed sample sets.

Once PT and AT subsets were defined, the DIMTEST procedure was run and a T-statistic was calculated using the following equation:

$$T = \frac{T_L - \bar{T}_B}{\sqrt{2}}, \tag{4}$$

where T_L represents the sum of the conditional covariances between AT items for examinees with the same score on PT items, and $\bar{T}_B$ is a bias correction term similar to T_L, calculated using simulated data (Walker, Azen, Schmitt, 2006). The T_L statistic is calculated as follows:

$$T_L = \frac{\sum_{k=1}^{K} T_{L,k}}{\sqrt{\sum_{k=1}^{K} S_k^2}}, \tag{5}$$

where k denotes all examinees, who received the same score on PT items, $T_{L,k}$ represents the conditional covariances, and S_k^2 equals the asymptotic variance of $T_{L,k}$. More specifically, $T_{L,k}$ can be defined as:

$$T_{L,k} = \frac{\sigma_{X,k}^2 - \sum_{i=1}^{N_1} p_{i,k} q_{i,k}}{2} = \sum_{i<l,k} Cov(U_{i,k}, U_{l,k}), \tag{6}$$

where *i* and *l* represent a given item pair with examinee scores on the given items represented by $U_{i,k}$ and $U_{l,k}$ respectively. For an overview of the DIMTEST procedure in greater detail, I refer the readers to Stout et al. (1993) and Nandakumar and Stout (1993). Additionally, while DIMTEST is originally a part of the nonparametric, conditional covariance-based dimensionality package, DIMPACK 1.0, it was implemented using the original code from the author of this study to bypass existing sample size restrictions in the GUI version of the software.

Additionally, exploratory DETECT results were used to define the factor structure for CFA. That is, the exploratory DETECT output providing a simple structure mapping of items to factors was used to specify the CFA, as well as the observed covariance matrix. While an in depth, technical overview of the elements of CFA is beyond the scope of this study, the basic model equation for CFA can be defined as:

$$\sum = \Lambda\Phi\Lambda' + \Theta_{\delta}, \tag{7}$$

where $\sum$ represents the population covariance matrix, Λ is the matrix of loadings specifying the relationship between observed variables and latent variables (consistently referred to as factor or dimensions throughout the current study), Φ is the covariance matrix for the latent variables, and Θ_{δ} represents the covariance matrix of measurement errors.

The CFA was completed across every iteration of each sample set using the *lavaan* package in R (Rosseel Y, 2012) with maximum likelihood estimation. Once the CFA and had been employed with the exploratory DETECT structure, the focus shifted to

the relevant model fit statistics. Specifically, the root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), and comparative fit index (CFI) were used to evaluate the performance of the CFA model across sample sizes.

After analyses were completed, results from the confirmatory DIMTEST and CFA models were compared across bootstrapped samples to determine if the confirmatory results consistently support exploratory factor structure obtained from the prior DETECT analysis. Additionally, the means and variances of the DIMTEST *T*-statistics were analyzed to determine if there is a sample size at which the baseline factor structure starts to become less dimensionally distinct, as well as to further inform the stability conclusions regarding the DIMTEST procedure. Similarly, means and variances of relevant fit statistics obtained from the CFA were likewise compared across sample sizes. While a direct comparison of the respective statistics from the employed confirmatory techniques would not be appropriate, relative interpretations provided a general framework for comparison of performance across methods.

Results

Data

As previously noted, the original dataset consisted of responses from 25,000 examinees on 60 items. The mean total score from the original dataset was 30.87 (*sd=11.65*), indicating that the average student answered just over 50% of the questions correct. Additionally, the median total score was 30, slightly lower than the mean, and visual inspection of the density distribution of total scores indicated that the data is very slightly positively skewed. Consistency in sampling distributions was also inspected, with

samples increasing in consistency as sample size increases, but remaining within reasonable deviation at the lower sample sizes as well. Figure 6.1 provides the sample consistency results.

Exploratory Results and Factor Difficulty

Results from previous research employing the exploratory DETECT analysis indicated that there were 15 factors in the original dataset. The minimum factor size was two items and the maximum factor size was 7 items (Table 6.1). Most factors (50%) consisted of three items. Factor difficultly, defined in this study as the mean proportion correct across items in the factor, ranged from .24 to .84 with a median value of .48 (Table 6.1).

Factor Correlations

Factor correlations were computed for the original set of data with 8% of factor pairs having a weak correlation (0.3-0.5), 34% having a moderate correlation (0.5-0.7), 52% having a high correlation (0.7-0.9), and 8% having a very high correlation (>0.9). Factor correlations from the original data can be found in Table 6.2. The mean correlation across all factors is .711 (*sd=.148*).

Additionally, factor correlations were computed for each sample across the five sample sizes, however, not all samples resulted in a reliable set of factor correlations. That is, at the lower sample sizes (*n=250; n=500*) the majority of samples resulted in one of two errors from *lavaan* (Rosseel Y, 2012): (1) That the covariance matrix of latent variables was not positive definite, or (2) That the information matrix could not be inverted and the standard errors could not be computed. Due to these errors the

corresponding correlation matrices contained NA (i.e., missing) values and/or values greater than 1. In these instances, the full correlation matrix from the affected sample was removed from the analysis of correlations. For samples of size 250, just 3 samples resulted in a reliable correlation matrix, while samples of size 500 saw only a very slight improvement of 7 matrices, samples of size 2,500 significantly improving with 42 matrices, and the two higher sample sizes retaining all 50 correlation matrices.

Despite losing a lot of information from the small samples, the correlation matrices remained relatively stable across all five samples sizes. The minimum mean correlation across all factor pairs and available matrices occurred at the sample size of 250 with a correlation coefficient of .68 (*sd=.17*), while the maximum mean correlation across all factor pairs occurred at the sample size of 25,000 with a value of .71 (*sd=.14*). See Figures 6.2 and 6.3 for full correlation results.

CFA Fit

As anticipated, all measures of fit improved and grew more stable as sample size increased. The minimum mean CFI value was 0.83, indicating poor fit at a sample size of 250, while the maximum mean CFI value was 0.95, indicating marginal fit at a sample size of 25,000. The maximum mean RMSEA value was .03 at a sample size of 250 and the minimum mean value was .02 at a sample size of 25,000, mean RMSEA values across all sample sizes indicated good fit (MacCallum, Browne, & Sugawara, 1996). Finally, the maximum mean SRMR value was .05 at a sample size of 250 and the minimum mean value was .02 at a sample size of 25,000, mean SRMR values across all

sample sizes indicated good fit (Hu & Bentler, 1999). Full CFA fit results are provided in Figure 6.4.

DIMTEST Statistics

Results of the DIMTEST analysis indicated that as sample size increased, a greater proportion of factor pairs became statistically dimensionally distinct from one another. More specifically, at a sample size of 250, only 5% of factor pairs were dimensionally distinct from one another while at a sample size of 25,000 the percentage of dimensionally distinct factor pairs increased to 93%. The maximum mean *p*-value was .267 (*sd=.113*) at a sample size of 250, indicating that on average, the factor pairs were not dimensionally distinct from one another. The minimum mean p-value was .032 (*sd=.142*) at a sample size of 25,000, indicating that on average, the factor pairs were dimensionally distinct. Results of the DIMTEST statistics can be found in Figures 6.5 and 6.6.

Discussion

Main Findings

While it isn't uncommon to pair DETECT and DIMTEST analyses, no evidence could be found of pairing exploratory DETECT results with a CFA. As such, this study was not approached with expectations for the overarching results, but rather as a simple exploration. The DIMTEST results appeared to provide strong support for the DETECT-defined factors, with support increasing, in the form of decreasing p-values, as sample size increased. The CFA fit also increased as sample size increased, signaling growing support for the DETECT-defined factors similar to the DIMTEST results. Drawing any

concrete conclusions regarding which method provides more support for the DETECT factors using just the DIMTEST p-values and CFA model fit indices is extremely difficult as they are not equivalent measures that can be directly compared, however, the overabundance of significant p-values (93%) at the sample size of 25,000 seems extremely compelling, while the marginal to good fit indicating by the CFA fit statistics seems less so. It is necessary to note that this is just a surface level interpretation and that much more research is needed to indicate which method should be given more weight.

Discriminant Validity Evidence

While the *p*-values provided by the DIMTEST procedure suggests that strong discriminant validity evidence for the DETECT factors may be present, the CFA factor correlations tell a slightly different story. More specifically, while only 7 factor pairs were not dimensionally distinct at the sample size 25,000 according to the DIMTEST analysis, 22 factor pairs had correlations greater than .85 via the CFA. With correlations that high, it is difficult to claim discriminant validity evidence of those 22 factors, which equates to about 21% of the total factor pairs.

These results reveal two main issues in drawing concrete conclusions regarding discriminant validity evidence within the current study. The first, and more pertinent issue, is that guidelines in determining which analysis to put greater weight on currently do not exist. Further exploration, such as conducting a simulation study, is needed to determine which analysis should be regarded as more reliable for indicating discriminant validity evidence. Additionally, it must be acknowledged that evaluating discriminant validity evidence in the context of this study shines a light on the tension being drawn

between substantive and statistical dimensionality. The real data used in this study comes from a single math test, in which, it may be reasonable to assume that any factors existing within it are highly correlated and would not exhibit discriminant validity. However, from a statistical perspective, this study *may* provide evidence to the contrary, and as such, should be further investigated in future research.

Limitations

Despite the insightful results obtained from the current study, it was not without its limitations. One major limitation was the number of items comprising some of the factors. That is, when using the DIMTEST procedure, Deng et al. (2012) recommend that the AT subset be comprised of *at least* four items, with the PT subset being larger than the AT subset (Deng et al., 2012). Of all the factor pairs analyzed for dimensional distinctness in this study, only 18 (17%) met both the AT and PT subset criteria. Additionally, only seven factors (47%) met the minimum number of items criteria for analysis.

Similarly, the number of items per factor may have been insufficient for the CFA, particularly within some of the lower sample sizes included in the study. While a wide variety of recommendations exist in the literature regarding sample size, number of factors, and the item to factor ratios necessary for adequate use of CFA, there is no one rule of thumb to adhere to. One of the many explorations in the literature includes Mundfrom, Shaw, and Ke (2005) who conducted a simulation study evaluating the relationship between sample size, number of factors, and item to factor ratios to determine thresholds for maintaining good recovery of the population solution. Results of the study by Mundfrom, Shaw, and Ke (2005) indicated that with 7 items per factor, the

necessity for larger sample sizes decreases, while ratios of fewer items per factor may necessitate thousands of examinees to adequate recover to population solution. By these standards only 2 factors (13%) met the suggested minimum item to factor ratio for result stabilization.

Future Directions

While the current study demonstrated the applicability of confirmatory methods to exploratory results with an increase in support of the exploratory factor solution as sample size increased, additional research is warranted to determine and parse out the effect due solely to an increase in data points. That is, a follow-up analysis should be conducted to address the question: when compared to a series of random factor models using the same data, do the results of the current study demonstrate a significant improvement in model fit? More specifically, it would be beneficial to create a null distribution using the data from the current study to proceed with another variation of formal hypothesis testing, the null vs. the current exploratory factor solution.

References

American Educational Research Association, American Psychological Association, Joint Committee on Standards for Educational, Psychological Testing (US), & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bosscher, R. J., & Smit, J. H. (1998). Confirmatory factor analysis of the general self efficacy scale. *Behaviour research and therapy*, *36*(3), 339-343.

Deng, N., Han, K. T., & Hambleton, R. K. (2013). A Review of DIMPACK Version 1.0: Conditional Covariance–Based Test Dimensionality Analysis Package. *Applied Psychological Measurement*, *37*(2), 162-172.

Finch, H., & Habing, B. (2007). Performance of DIMTEST-and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement*, *31*(4), 292-307.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Jang, E. E., & Roussos, L. (2007). An Investigation into the Dimensionality of TOEFL Using Conditional Covariance-Based Nonparametric Approach. *Journal of Educational Measurement*, *44*(1), 1-21.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data (Doctoral dissertation, University of Illinois at Urbana-Champaign). Dissertation Abstracts International: Section, B, 5598, 12-55.

Lay, A. L. (2020a). *A Comparative Study of the Accuracy and Precision of Exploratory Dimensionality Analyses* (doctoral dissertation excerpt, The University of North Carolina at Greensboro).

Lay, A. L. (2020a). *Evaluating the Stability of Factor Definitions in the Exploratory Detect Procedure* (Doctoral dissertation excerpt, The University of North Carolina at Greensboro).

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, *1*(2), 130.

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport*, *60*(1), 48-58.

McCrae, R. R., Zonderman, A. B., Costa Jr, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, *70*(3), 552.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Mueller, R. M., Lambert, M. J., & Burlingame, G. M. (1998). Construct validity of the Outcome Questionnaire: A confirmatory factor analysis. *Journal of Personality Assessment*, *70*(2), 248-262.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159-168.

Nandakumar, R., & Yu, F. (1996). Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of educational measurement*, *33*(3), 355-368.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses-comparison of different approaches. *Journal of educational measurement*, *31*(1), 17-35.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of educational statistics*, *18*(1), 41-68.

Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, 48(2), 1–36. http://www.jstatsoft.org/v48/i02/.

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, *43*(3), 215-243.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical factor analysis to detect multidimensionality. *Journal of educational measurement*, *35*(1), 1-30.

Seraphine, A. E. (2000). The performance of DIMTEST when latent trait and item difficulty distributions differ. *Applied Psychological Measurement*, *24*(1), 82-94.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589-617.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). DIMTEST manual. *Unpublished manuscript available from WF Stout, University of Illinois at Urbana-Champaign, Champaign*.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*(4), 331-354.

Svetina, D., & Levy, R. (2012). An overview of software for conducting dimensionality assessment in multidimensional models. *Applied Psychological Measurement*, *36*(8), 659-669

Walker, C. M., Azen, R., & Schmitt, T. (2006). Statistical versus substantive dimensionality: The effect of distributional differences on dimensionality assessment using DIMTEST. *Educational and Psychological Measurement*, *66*(5), 721-738.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika, 64, 213-249.

Tables

Table 6.1

Factor Size and Difficulty

Factor	N Items	Difficulty
1	7	.68
2	3	.84
3	7	.60
4	3	.72
5	4	.65
6	3	.65
7	6	.57
8	3	.48
9	3	.41
10	4	.31
11	5	.35
12	2	.41
13	3	.26
14	3	.33
15	4	.24

Table 6.2

DETECT Defined Factor Correlations from Lavaan

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
F1	1	0.795	0.769	0.918	0.892	0.906	0.927	0.802	0.690	0.559	0.747	0.713	0.500	0.523	0.594
F2	0.795	1	0.581	0.693	0.725	0.649	0.616	0.506	0.468	0.333	0.461	0.441	0.324	0.330	0.355
F3	0.769	0.581	1	0.907	0.897	0.755	0.761	0.815	0.803	0.687	0.879	0.726	0.612	0.560	0.663
F4	0.918	0.693	0.907	1	0.945	0.901	0.893	0.827	0.780	0.600	0.84	0.727	0.549	0.550	0.656
F5	0.892	0.725	0.897	0.945	1	0.889	0.892	0.908	0.835	0.715	0.881	0.780	0.629	0.570	0.686
F6	0.906	0.649	0.755	0.901	0.889	1	0.931	0.819	0.732	0.624	0.765	0.722	0.519	0.499	0.634
F7	0.927	0.616	0.761	0.893	0.892	0.931	1	0.880	0.750	0.653	0.816	0.793	0.559	0.561	0.641
F8	0.802	0.506	0.815	0.827	0.908	0.819	0.880	1	0.880	0.742	0.883	0.815	0.642	0.720	0.722
F9	0.690	0.468	0.803	0.780	0.835	0.732	0.750	0.880	1	0.735	0.858	0.731	0.741	0.570	0.697
F10	0.559	0.333	0.687	0.600	0.715	0.624	0.653	0.742	0.735	1	0.804	0.862	0.720	0.501	0.755
F11	0.747	0.461	0.879	0.840	0.881	0.765	0.816	0.883	0.858	0.804	1	0.880	0.841	0.650	0.860
F12	0.713	0.441	0.726	0.727	0.780	0.722	0.793	0.815	0.731	0.862	0.88	1	0.722	0.613	0.808
F13	0.500	0.324	0.612	0.549	0.629	0.519	0.559	0.642	0.741	0.720	0.841	0.722	1	0.588	0.839
F14	0.523	0.330	0.560	0.550	0.570	0.499	0.561	0.720	0.570	0.501	0.650	0.613	0.588	1	0.622
F15	0.594	0.355	0.663	0.656	0.686	0.634	0.641	0.722	0.697	0.755	0.860	0.808	0.839	0.622	1

Figures

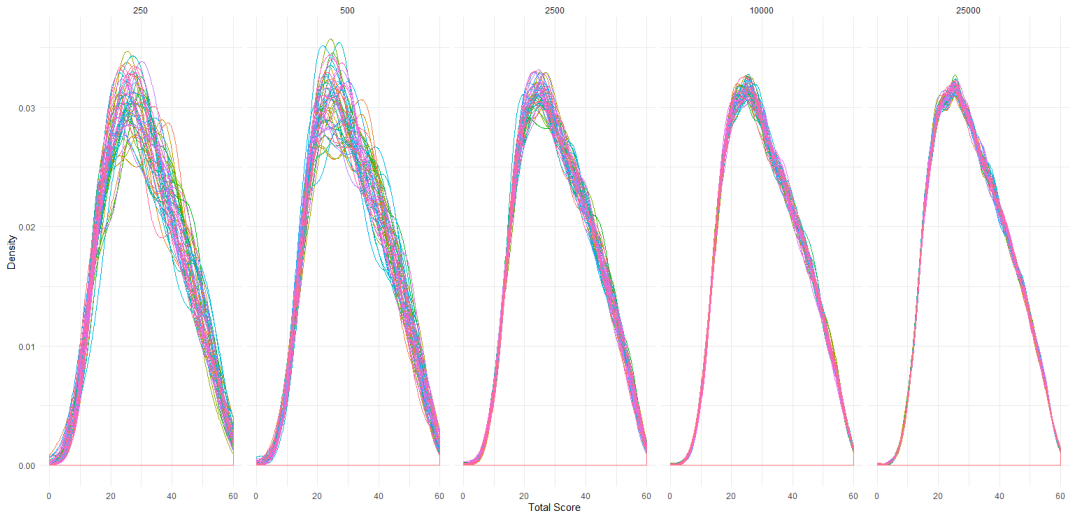

Figure 6.1. Overlaid Density Distributions of Total Score by Sample Size.

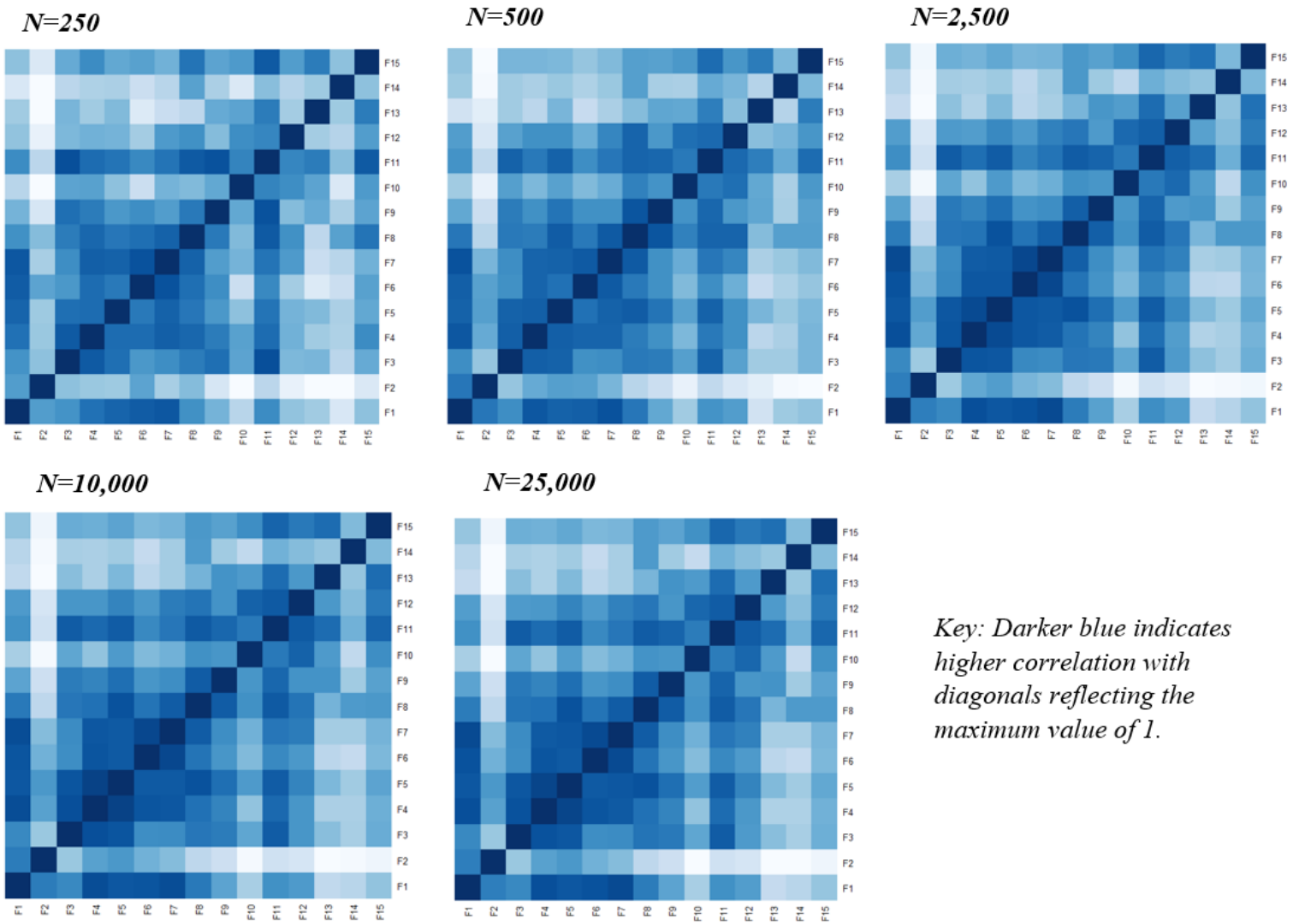

Figure 6.2. Mean Factor Correlations by Sample Size.

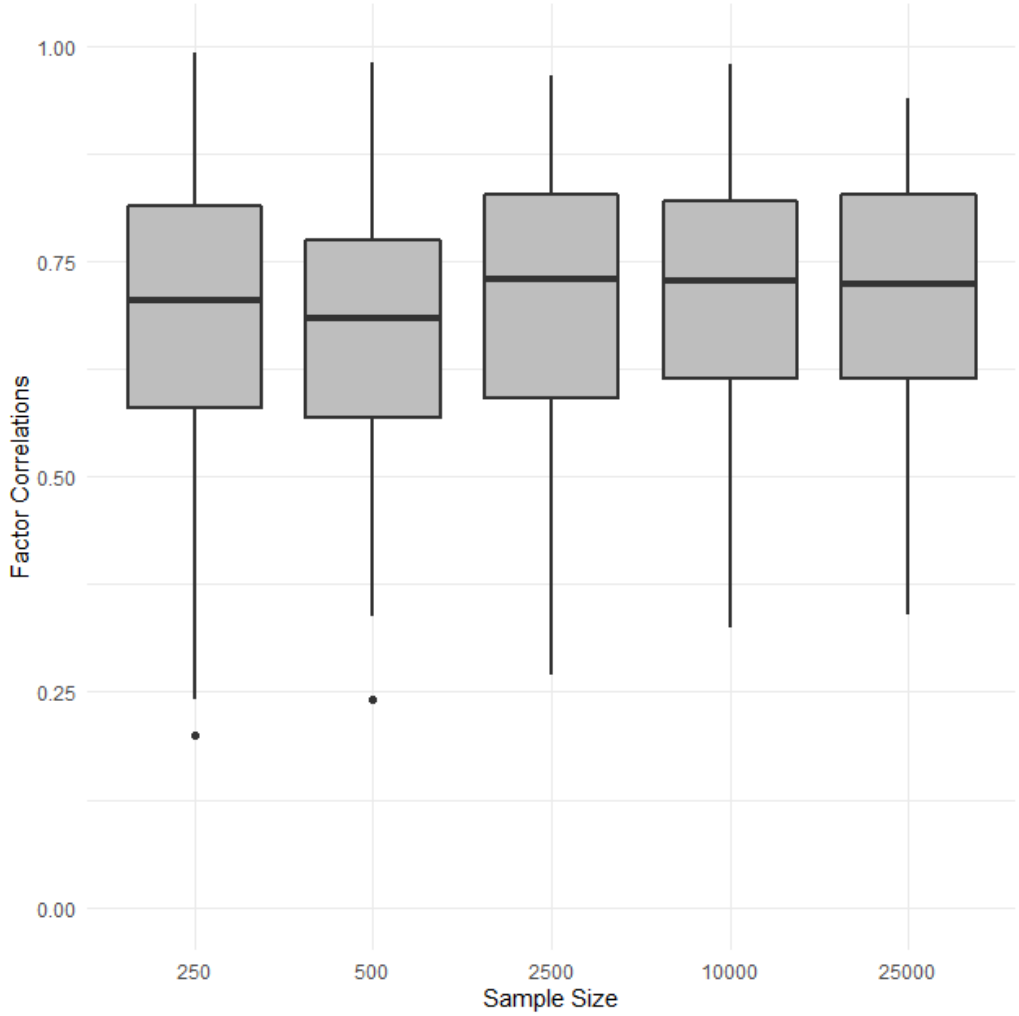

Figure 6.3. Factor Correlations by Sample Size.

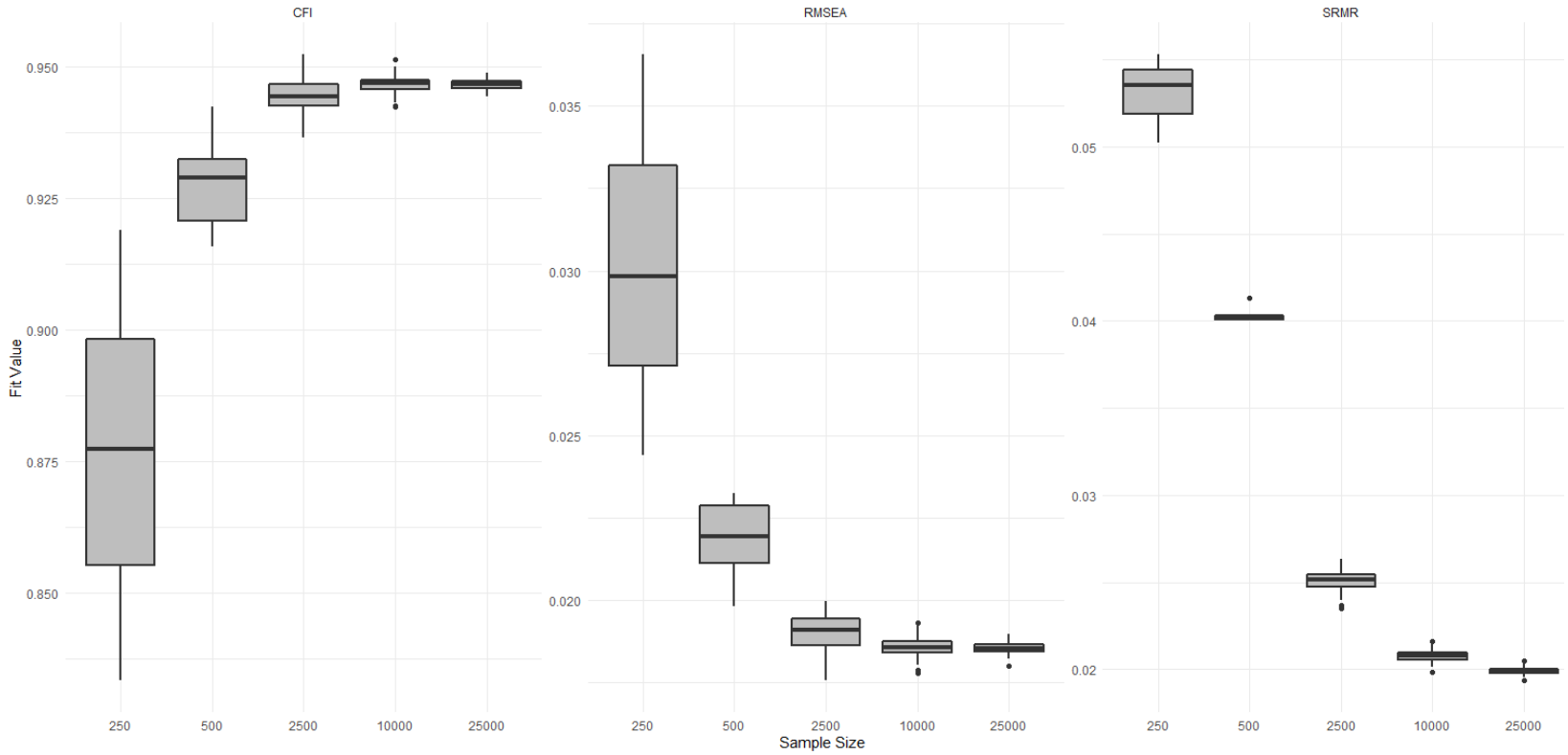

Figure 6.4. CFA Fit Statistics by Sample Size.

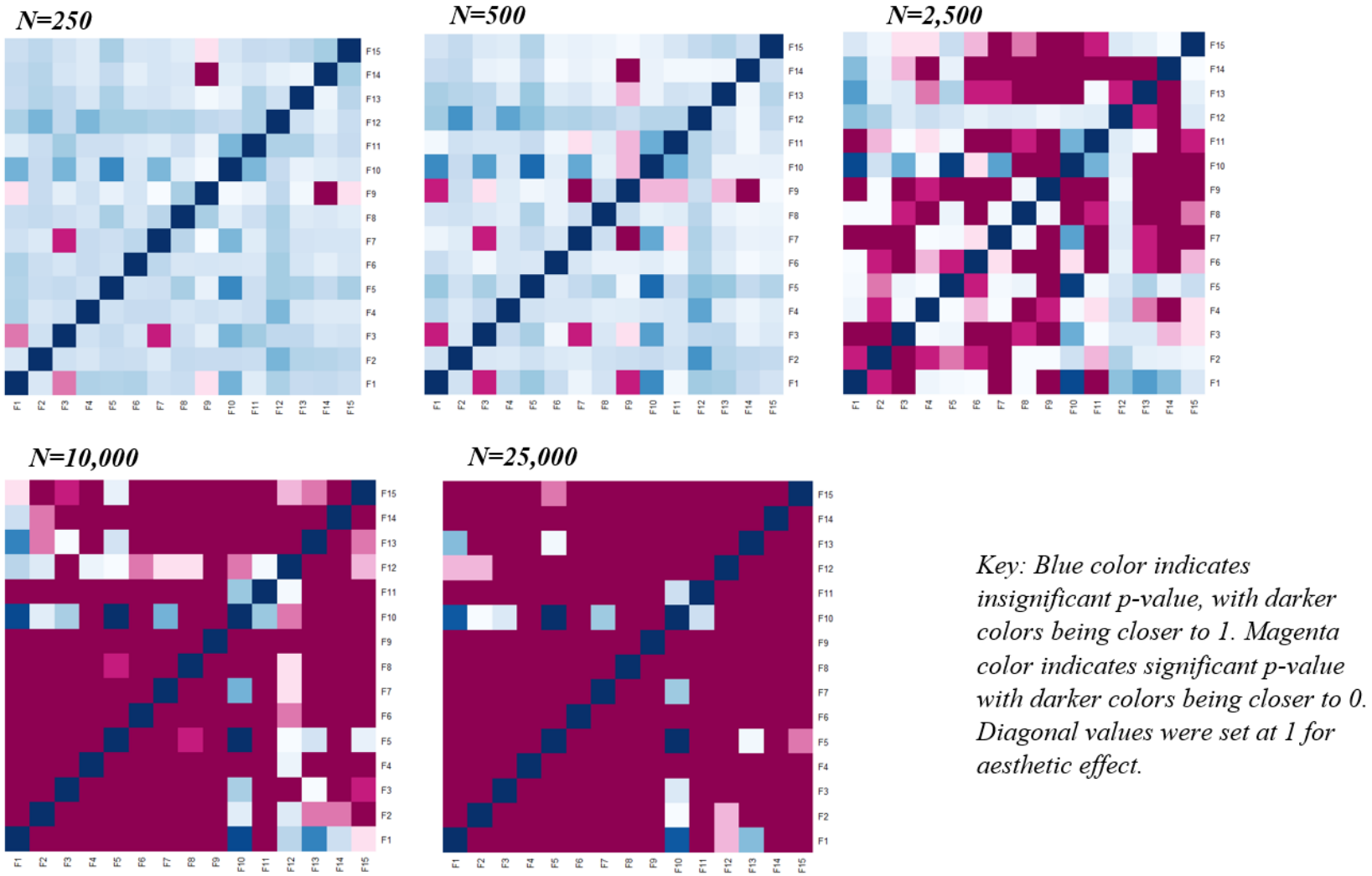

Figure 6.5. Mean DIMTEST P-values by Factor and Sample Size.

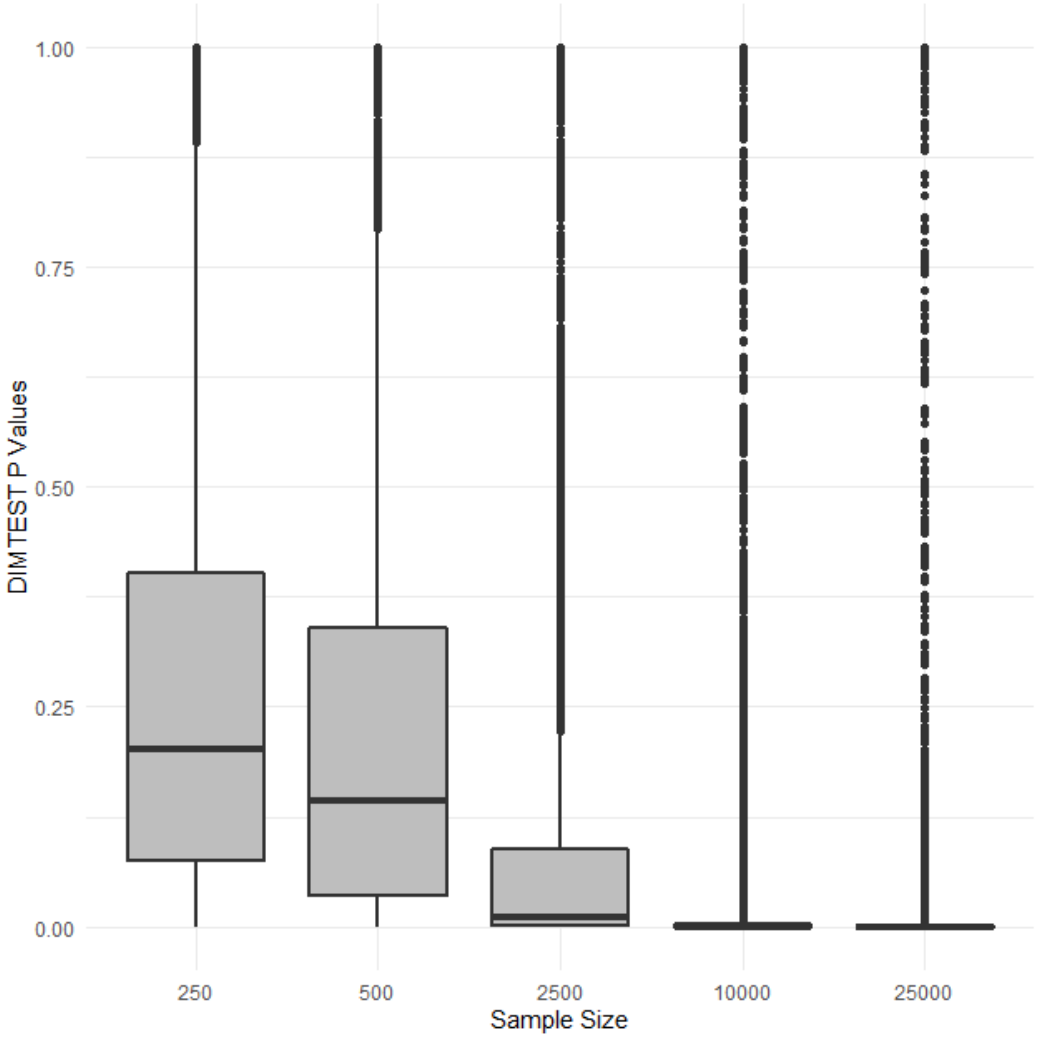

Figure 6.6. DIMTEST P-values by Sample Size.

CHAPTER VII

FINAL DISCUSSION

Overview of Purpose and Main Findings

The goal of this dissertation was to evaluate the performance of various exploratory and confirmatory dimensionality analyses across five different sample sizes when applied to real student assessment data. The first study aimed to evaluate the accuracy and precision of three exploratory dimensionality analyses: the DETECT procedure, a parallel analysis implementing principle components analysis, and a parallel analysis implementing common factor analysis. Results of the study indicated that the PCA was the most accurate and precise method of the three, with DETECT being the least precise and systematically under extracting the number of factors relative to the "true" (i.e., full sample) solution, and common FA varying in degrees of precision and accuracy across sample sizes, moving from under extraction to over extraction. However, the study did not result in any sample size-based recommendation, instead it emphasized the necessity for researchers to consider the purpose of the study when determining which dimensionality analysis to utilize, specifically underscoring the difference between methods meant to model the latent structure in the data and those intended purely for data reduction.

The second study served as an extension of the first, evaluating the accuracy and stability of the DETECT procedure at a finer grain size. In this study the factor solution,

or item to factor mapping, served as the basis for comparison. Generally, the second study aimed to explore whether evaluating DETECT at the factor solution level provides any additional information regarding the accuracy and stability of the analysis above and beyond the results provided in the first study. DETECT factor congruence values were used to evaluate accuracy at the factor solution level and indicated that despite the number of clusters leveling out at a given sample size, as demonstrated in the first study, the overlap between sample results and relative truth continues to increase with sample size. Overall, while evaluating the performance of the DETECT procedure using the more parsimonious approach may be adequate for most use cases of the analysis, the more fine-grained evaluation does provide a greater degree of accuracy.

The final study shifted away from exploratory dimensionality analyses to evaluate the implementation of confirmatory analyses with exploratory results. As with the second study, the third served as an extension of the previous research conducted in this dissertation, using the exploratory factor solution from papers one and two to apply confirmatory analyses. Results indicated that the DIMTEST procedure provided good support for the exploratory DETECT factors, specifically at larger sample sizes. The results of the CFA were slightly mixed, there was an increase in model fit by sample size indicating good support for the DETECT factors, however, the analysis also resulted in a relatively high proportion of high factor correlations Given these results, it is difficult to draw any concrete conclusions regarding discriminant validity evidence. That is, additional research, such as a simulation study, is needed in order to better flesh out which method is more reliable for providing discriminant validity evidence.

Limitations

While employing the analyses included in the three studies with real data rather than through simulation was originally outlined as (and remains) a major benefit of the studies, it also introduces some limitations. Specifically, using response data from a single form of a single assessment results in a very specific context in which the results can be directly interpreted. This means that while each study provided insightful results, the inclusion of additional forms of the current assessment and additional assessments/assessment contexts are necessary in order to be able to adequately generalize the results to a larger context or help flesh out any potential inconsistencies across forms and further evaluate *why* they may exist.

Another limitation that arose was the maximum sample size. Since the samples were composed from a real dataset, they could not exceed that original sample size of 25,000. While typically 25,000 examinees are considered more than enough when evaluating dimensionality, there appeared to be room for continued improvement in some of the measures utilized throughout the studies. For example, both the ratio R and FDC values continued to increase with sample size, but both still may have improved more had there been additional datapoints to include in analysis.

Moreover, the number of factors determined by the exploratory DETECT procedure and subsequently used throughout all studies was very large. While this dissertation did not focus heavily on the substantive perspective of dimensionality, it should be acknowledged that the it would be extremely difficult to determine the substantive meaning of 15 factors across 60 items. Moreover, it posed serious issues

when evaluating dimensionality from a confirmatory perspective, in some cases failing to meet the minimum requirements for item-to-factor ratios for both confirmatory methods utilized.

Methodological Considerations for Future Research

As noted within the limitations, all three studies suffer from a lack of generalizability, and as such would benefit from future research employing additional assessment contexts. There are a wide variety of avenues in which this research could progress down. For one, it would be beneficial to replicate the current studies using multiple forms of a single assessment to test whether results are comparable within assessment, across forms. Additionally, replication across diverse assessment contexts is necessary to be able to extrapolate the resulting inferences outside of their current assessment-specific state. As repeatedly specified, dimensionality is a product of the interaction between the student and the assessment. As such, it is reasonable to assume that that interaction would look fundamentally different not only between assessments, but also between assessment types. The current studies are housed with the standardized assessment context, future research should aim to expand that purview to more fluid, dynamic assessments, such as game-based assessments, in which the type and degree of interaction occurring differs drastically.

Another consideration for future research is an exploration of the maximum DETECT value and its relationship to DFC. That is, results of the second study indicated that both the ratio R and FDC values increase with sample size. However, the DETECT value levels off after a sample size of 2,500, implying that the maximum DETECT value

is actually decreasing across samples of size 2,500 to 25,000. To make plain the interpretive discourse at present: the combination of these results indicates that the degree of multidimensionality present in the data levels out after samples of size 2,500, the maximum potential degree of multidimensionality appears to decrease across the same set of samples, and the overlap between factor solutions continues to increase. Further research is needed to understand, in detail, the relationship these DETECT-related measures have to each other across sample sizes.

A final methodological consideration for future research that emerged not from the results of one of the three studies, but instead as part of the process building the necessary code to run all of the analyses, is a closer inspection of the performance of the simulation technique utilized in DIMTEST. While not talked about in detail in this dissertation, it is mentioned that the DIMTEST procedure uses a non-parametric IRT model to evaluate the data and subsequently simulate additional samples for the creation of the T_L statistic. The process of writing out this analysis called into question the payoff of its complexity. While typically non-parametric methods are celebrated for their lack of computational complexity, in this particular case it seemed as though fitting a simple Rasch model to simulate data from might actually be less complex. Future studies should consider comparing alternative methods for data generation such as sampling with replacement or using a Rasch model.

Substantive and Interactive Considerations for Future Research

Lastly, and most importantly, although the focus of this dissertation was on evaluating and comparing the statistical perspective of dimensionality, it would be remiss

to ignore the substantive stake in this game. There are three main substantive considerations that need to be given adequate attention in future endeavors of dimensionality research. The first is considerations of the students themselves. What characteristics do these students possess, and do those characteristics affect the dimensionality of an assessment? An abundance of research, some cited in this dissertation, suggests that student characteristics play a large role in determining statistical dimensionality, yet this seems to be an area that gets very little attention in recent work on assessing dimensionality.

The second are for consideration are the more traditional notions of substantive characteristics of the assessment. That is, content area, test structure (e.g., testlet based assessments), cognitive load, should all be considered when evaluating dimensionality. While exploratory analyses allow us to more easily pick up shifts in statistical dimensionality, if those results do not make reasonable sense in terms of substantive characteristics of the test, they become extremely difficult, and often times impossible, to adequately interpret. Instead researchers may consider starting with the substantive aspects to build out substantively supported dimensions to be tested in a confirmatory context.

Finally, and most importantly, the field of educational and psychological dimensionality research *must* give greater consideration to the types of interactions occurring within an assessment context and the complex stimuli that elicit those interactions. Assessments continue to get more innovative and interactive. If the shifting paradigm in the way students are regularly interacting with assessments is not taken

seriously, dimensionality research will continue to be restricted to exploratory analyses in the more fluid contexts and productive research on effectively analyzing dimensionality of these complex assessment will become stagnant. Instead a collaborative effort should be made with educational psychology experts to study and categorize these types of interactions in order to start to build a more sustainable foundation for defining substantive factors within dynamic assessments that may eventually be evaluated from a confirmatory perspective.